AFRL-AFOSR-VA-TR-2017-0020

Dynamic Generalizations of Systems Factorial Technology for Modeling Perception of Fused Information

Joseph Houpt
WRIGHT STATE UNIVERSITY
3640 COLONEL GLENN HIGHWAY
DAYTON, OH 45435-0001

01/11/2017
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/RTA2

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to the Department of Defense, Executive Service Directorate (0704-0188). Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-11-2016	Final	01-03-2013 -- 31-08-2016

4. TITLE AND SUBTITLE
Dynamic Generalizations of Systems Factorial Technology for Modeling Perception of Fused Information

5a. CONTRACT NUMBER
FA9550-13-1-0087

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Houpt, Joseph W.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Wright State University
3640 Colonel Glenn Hwy
Dayton, OH 45435

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Office of Scientific Research
875 North Randolph Street
Suite 325, Room 3112
Arlington VA, 22203

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION A: Distribution approved for public release.

13. SUPPLEMENTARY NOTES

14. ABSTRACT
Models are a fundamental part of understanding cognition. The advantages of cognitive modeling are particularly clear when attempting to understand how changes in a cognitive task lead to changes in performance. Systems factorial technology (SFT) can be used to explain and understand why there are differences in performance, not just that there is a difference. In this project, we have extended the applicability of SFT to more complex environments than the basic perceptual experiments to which it has been previously applied. This included extensions of the statistical analyses to include hierarchical parametric Bayesian modeling and semi- and non-parametric modeling. We then applied SFT in both basic visual search studies and in task requiring the use of multispectral imagery.

15. SUBJECT TERMS
cognitive modeling; Bayesian statistics; visual search; multispectral imagery; image fusion

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Joseph W. Houpt
U	U	U	UU	100	19b. TELEPHONE NUMBER *(Include area code)* (812) 202-2509

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18
Adobe Professional 7.0

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

1 Overview

Models are a fundamental part of understanding cognition. The advantages of cognitive modeling are particularly clear when attempting to understand how changes in a cognitive task lead to changes in performance. Standard statistical tools can answer simple comparative questions, such as whether people perform better when using one display type relative to another. Standard statistics do not provide insight into the mechanism that lead to the change in performance. Cognitive models can be used to explain and understand *why* there are differences in performance, not just that there is a difference. Many, if not most, cognitive tasks involve the combination of multiple sources of information to make a decision. These sources can include information across multiple modalities, such as using a person's voice and a facial image to identify him or her; or there can be multiple sources of information within a single modality, such as the use of the shape and color of an object to determine if it is a weapon. In many cases the sources are based on the experimental design and not necessarily on psychologically meaningful features, such as the top half and bottom half of a face (Burns, Pei, Houpt, & Townsend, 2009).

Systems Factorial Technology (SFT; Townsend & Ashby, 1983a; Townsend & Nozawa, 1995a) is a framework for studying how different sources of information combine in cognitive processing. These sources can be as similar as visual information from the left and right visual field or as disparate as the demands of two different tasks such as driving while talking on a cell phone. SFT stands out as particularly powerful framework because the various manners in which information can be combined are classified based on mathematically defined model properties. Despite the rigor of the definitions, SFT is quite general in that it requires no distributional or parametric assumptions about the cognitive processes. Using these precise mathematical definitions, there are multiple tests within the SFT framework to reject large classes of possible processing properties and support very specific properties. These properties can be grouped into four categories: Architecture, stopping-rule, stochastic dependence and workload capacity, each of which will be defined below (Townsend, 1974).

Information visualization is a domain in which multiple sources of information are often combined in various ways to optimize the human perception of that information. By using SFT to study the display of visual information, we can gain a better understanding of the processes underlying a person's perception of the information. This understanding can be used to target specific improvements in the way information is presented and combined in a display. This project is concerned with a particular type of visualization, the display of information from multiple spectra, particularly visual and long-wave infrared light.

Extensive literature is available on image fusion, including on the combination of visual and infrared imagery. One of the main goals of the field is to find ways to combine these sources in the "best" way possible. Often "best" is determined based on the information content (in the information theory sense) of the combined image, rather than directly based on human performance using

DISTRIBUTION A: Distribution approved for public release.

the combined image. While information theoretic measures are often a good proxy for measures of human performance (e.g., Fitts Law; Fitts, 1954), they are indirect and do not give any account of the underlying processes. SFT offers an ideal framework for both assessing information fusion algorithms and for understanding why people do better or worse when using imagery fused with the various algorithms.

In its original form, the SFT framework was best suited for use with static stimuli, e.g. photographs, rather than dynamic stimuli, e.g. video. While there are a number of important applications of the fusion of static images, dynamic imagery is often available and more informative. As part of this project, we generalized SFT for use with complex, dynamic imagery. To achieve that goal, we extended the inferential statistical tools available for SFT and examined dynamic response measures with the SFT framework, particularly eye-tracking and mouse-tracking. SFT is a general methodology, not specific to information visualization, so this generalization will have a much broader impact on the study of cognitive processing. The integration of SFT with eye- and mouse-tracking will also be beneficial for researchers who collect those types of data because it will go beyond traditional statistical analyses to offer an account of the processes that lead to those data.

2 Introduction

SFT provides a powerful and rigorous framework for studies of the perceptual and cognitive processes involved in combining multiple sources of information. The main aspects of combining information in cognitive processing within the SFT framework are divided into four classes: *architecture*, *stopping-rule*, *stochastic dependence* and *workload capacity*. *Architecture* refers to the temporal organization of the process, in particular whether the processes occur serially (e.g., first perceive then decide), in parallel (e.g., audio and visual information is processed simultaneously), or otherwise. Note that this use of architecture is not the same use that refers to fixed properties or parameters of an information processing system (e.g. in the ACT-R, SOAR, etc. frameworks). The *stopping-rule* determines how many of the sources must be processed before finishing or responding. For example, if a spoken word can be identified aurally, an individual may respond without waiting to process the information given by the shapes of the speaker's mouth as the word is spoken. *Stochastic dependence* refers to how the processing time of each source of information relates to the others and indicates interactions among the processes. *Workload capacity* refers to how the speed of processing a source of information changes as more sources are added.

The two main measures associated with SFT are the survivor interaction contrast (SIC) and the capacity coefficient. The SIC, developed originally by Townsend and Nozawa (1995a) and extended by many others (Dzhafarov, Schweickert, & Sung, 2004; Eidels, Houpt, Pei, Altieri, & Townsend, 2011a; Fifić, Nosofsky, & Townsend, 2008; Houpt & Townsend, 2010a, 2011), is a contrast of the response time distributions as the speed of processing of each source is

DISTRIBUTION A: Distribution approved for public release.

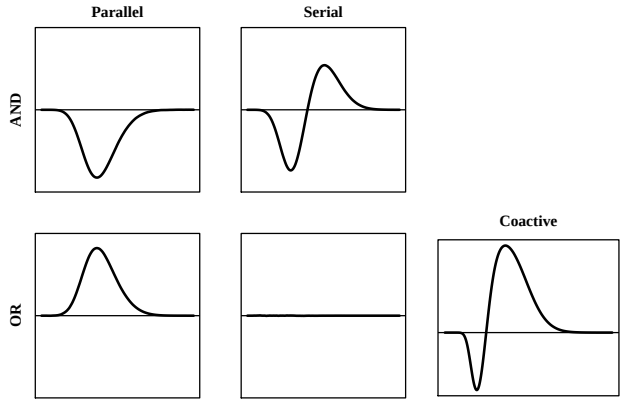

Figure 1: Survivor interaction contrast predictions for parallel and serial models with AND and OR stopping rules, assuming selective influence and for the Poisson and diffusion based coactive (information summing) models.

factorially manipulated. Using subscripts to indicate speed of processing, H for fast (high salience) and L for slow (low salience), with each position indicating the level on a source of information, and $S(t)$ indicating the survivor function (the probability that a response has not occurred by time t),

$$\mathrm{SIC}(t) = [\mathrm{S_{LL}}(t) - \mathrm{S_{LH}}(t)] - [\mathrm{S_{HL}}(t) - \mathrm{S_{HH}}(t)]\,. \tag{1}$$

Thus, the SIC is a functional measure, with values defined for each time point for which the response time distributions are defined (usually $t \in [0, \infty)$). This measure is particularly informative regarding the architecture and stopping-rule of a system. If the salience factors are chosen to selectively influence the processing of each source of information (the factor affects the processing time of targeted source and not the other), then each combination of parallel or serial processing with first-terminating and exhaustive stopping rules predicts unique SIC, shown in Figure 1. Parallel models with interactions that cause failures of selective influence also have relatively distinctive SIC shapes Eidels et al. (2011a). An extreme case of the interactive parallel model is the coactive model, in which the information from each source is pooled for a decision rather than treated separately. The coactive model has an SIC that is distinct from the models with selective influence, shown on the far right of Figure 1. Although the shape of the coactive and serial-and models appears similar, these models can be distinguished by testing the area under the SIC curve; the serial-and SIC has equal positive and negative areas leading to a net zero area under the curve, while the coactive model has a larger positive area under the curve.

$$\mathrm{MIC}(t) = [\mathrm{M_{LL}}(t) - \mathrm{M_{LH}}(t)] - [\mathrm{M_{HL}}(t) - \mathrm{M_{HH}}(t)]\,. \tag{2}$$

The area under the curve, or integrated, SIC is the mean interaction contract

DISTRIBUTION A: Distribution approved for public release.

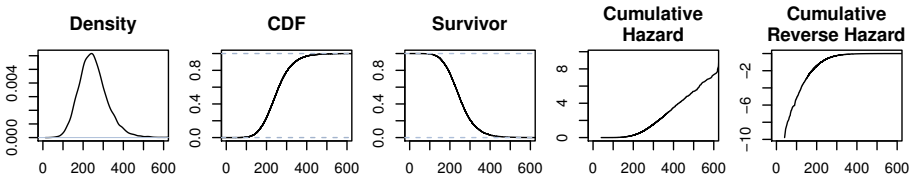

Figure 2: Different functions describing the same random variable. On the far left is the density (or PDF). Next is the cumulative distribution function, $\Pr\{X \leq t\}$, then the survivor function, $\Pr\{X > t\}$. The final two graphs are the cumulative hazard function, $H(t)$ and the cumulative reverse hazard function $K(t)$.

(MIC) , MIC $= \int_0^\infty \text{SIC}(t)\ dt$. The sign of the MIC (given some conditions, or the mean response time ordering) can be used to diagnose two of the fundamental properties in cognitive operations. When each subprocess is selectively influenced in a serial system, then the MIC will be zero (regardless of stopping rule), whereas in a serial system the MIC will be non-zero. Parallel, exhaustive processing leads to MIC < 0 and parallel, first-terminating processing leads to MIC > 0. Like the parallel, first-terminating processes, coactive processes will also lead to MIC > 0.

While the SIC has more diagnostic power, the MIC has some advantages over the SIC for diagnosing underlying mental architectures. First, fewer trials are needed to achieve a good estimate of the MIC because it is a single value, unlike the SIC which is an entire function. In practice this means that running a study using MIC could require fewer trials then a study using SIC. If there is little constraint on the number of trials that can be collected, SIC might be preferred (e.g., Townsend & Fifić, 2004). In many cases, conducting a large scale study involving a large number of stimulus trials per subject is not a realistic scenario. Research participants, are usually reluctant to participate in lengthy studies, and are more likely to drop out. Hence, long term studies can require significant financial compensation to recruit and retain participants. Additionally, subjects from particular populations are only available for study participation for a brief period of time. This can be due to limited mental capabilities and are not able to focus for a long period of time, or due to other constraints on their time. For example, autistic children (cf. Johnson, Blaha, Houpt, & Townsend, 2010), or air force pilots (cf. Schreiber, Stock, & Bennett, 2006) would only be available to serve as experimental participants for a limited number of trials. In such situations it is highly impractical to conduct repeated study sessions limiting a researcher to a relatively smaller number of response trials.

The other main measure, the capacity coefficient, is based on the comparison of performance with a single source of information to performance with multiple sources. The functions can indicate variations in workload capacity as

DISTRIBUTION A: Distribution approved for public release.

well as dependencies among source processing times. Although a capacity coefficient could be defined for any stopping rule, the most commonly used are the OR capacity coefficient Townsend and Nozawa (1995a) and the AND capacity coefficient Townsend and Wenger (2004a). Each capacity coefficient compares performance on a given task to an unlimited-capacity, independent, parallel (UCIP) model using the given stopping rule.

In an OR process, the probability that a response has not yet been made (the survivor function of the response times) is the probability that a target has not yet been detected on any source. If we write $S_{AB}(t)$ for the survivor function of response times when both A and B are present targets and $S_{A(B)}(t)$ for the survivor functions of the source completion times on A when the B target is present (and likewise for B in the presence of A), then,

$$S_{AB}(t) = S_{A(B)}(t) \times S_{B(A)}(t).$$

With the additional UCIP assumptions, the completion time distribution of A is unchanged regardless of whether B is present, $S_{A(B)}(t) = S_A(t)$ and likewise for B, $S_{B(A)}(t) = S_B(t)$. Hence, a UCIP model predicts that the survivor function when both targets are present is equal to the product of the survivor functions for each target in isolations,

$$S_{AB}(t) = S_A(t) \times S_B(t).$$

The argument holds more generally; the survivor function for any number of targets is the product of the survivor function for each of those targets in isolation,

$$S_{1\ldots n}(t) = \prod_{i=1}^{n} S_i(t).$$

For both statistical reasons (Houpt & Townsend, 2012, cf.) and interpretability Townsend and Ashby (1983a); Townsend and Nozawa (1995a), the OR capacity coefficient is defined using cumulative hazard functions ($H(t)$). Getting from survivor functions to cumulative hazard functions requires simply taking the natural logarithm, $\log[S(t)] = H(t)$. Thus, because $\log(xy) = \log(x) + \log(y)$,

$$H_{1\ldots n}(t) = \log\left[S_{1\ldots n}(t)\right] = \log\left[\prod_{i=1}^{n} S_i(t)\right] = \sum_{i=1}^{n} H_i(t). \tag{3}$$

The OR capacity coefficient is defined as a participant's actual performance when all sources are present ($\hat{H}_{1\ldots n}(t)$), to her predicted performance if the UCIP assumptions are satisfied,

$$C_{\mathrm{OR}}(t) = \frac{\hat{H}_{1\ldots n}(t)}{\sum_{i=1}^{n} \hat{H}_i(t)}. \tag{4}$$

The denominator is the estimated cumulative hazard function for the UCIP model based on the response times for each process in isolation, and the numerator is the actual performance.

DISTRIBUTION A: Distribution approved for public release.

In an AND process, the probability that a response was made (the cumulative distribution function of the response times) is the probability that a target has been detected on each source. If we write $F_{AB}(t)$ for the CDF of response times when both A and B are present targets, $F_{A(B)}(t)$ for the CDF of the source completion times on A when the B target is present (and likewise for B in the presence of A), then,

$$F_{AB}(t) = F_{A(B)}(t) \times F_{B(A)}(t).$$

The UCIP assumptions imply that the CDF of an individual target detection time does not change with respect the presence of the other source, $F_{A(B)}(t) = F_A(t)$ and likewise for B, $F_{B(A)}(t) = F_B(t)$. Hence,

$$F_{AB}(t) = F_A(t) \times F_B(t).$$

More generally,

$$F_{1\ldots n}(t) = \prod_{i=1}^{n} F_i(t).$$

Like the OR capacity coefficient, we take the natural logarithm of both sides to obtain capacity coefficient predictions. Because the AND model is in terms of the CDF rather than the survivor function, this operation results in a cumulative *reverse* hazard function, $\log[F(t)] = K(t)$. Hence, the UCIP prediction for an AND task is that,

$$H_{1\ldots n}(t) = \log\left[S_{1\ldots n}(t)\right] = \log\left[\prod_{i=1}^{n} S_i(t)\right] = \sum_{i=1}^{n} H_i(t).$$

The AND capacity coefficient is defined as a participant's actual performance when all sources are present ($\hat{K}_{1\ldots n}(t)$), to predicted performance assuming the UCIP assumptions are satisfied,

$$C_{\text{AND}}(t) = \frac{\sum_{i=1}^{n} \hat{K}_i(t)}{\hat{K}_{1\ldots n}(t)}. \tag{5}$$

The numerator is the estimated cumulative reverse hazard function for the UCIP model based on the response times for each process in isolation, and the denominator is the actual performance. The UCIP prediction is in the numerator for $C_{\text{AND}}(t)$ so that the interpretation of values relative to one is consistent with $C_{\text{OR}}(t)$. Relatively larger cumulative hazard functions indicate faster processing, while relatively larger cumulative reverse hazard functions indicate slower processing.

$C_(t) < 1$, or limited-capacity, implies worse performance than the UCIP model. Limited capacity indicates that either there are limited processing resources, there is inhibition among the subprocesses, or the items are not processed in parallel (e.g., the items may be processed serially).

$C_(t) > 1$, or super-capacity, implies better performance than the UCIP model. Super capacity indicates that either there are more processing resources

DISTRIBUTION A: Distribution approved for public release.

available per process when there are more processes, that there is facilitation among the subprocesses, or the items are not processed in parallel (e.g., the items may be processed coactively).

In principal, SFT can apply to a wide range of domains. However the methods have thus far been applied to relatively simple, static stimuli. The original application was to the detection of dim dots appearing either in the left visual field or the right visual field Townsend and Nozawa (1995a). Since then the methodology has been applied to simple audio-visual (e.g., Altieri & Townsend, 2011) and visual only (e.g., Donnelly, Cornes, & Menneer, 2012; Fifić & Townsend, 2010; Johnson et al., 2010) discrimination tasks, change detection (e.g., C.-T. Yang, 2011; C.-T. Yang, Hsu, Huang, & Yeh, 2011), categorization (e.g., Fifić et al., 2008; Little, Nosofsky, & Denton, 2011), and memory search (e.g., Townsend & Fifić, 2004). Nonetheless, in all of these applications the stimuli were highly controlled and presented in isolation with little or no extraneous information. Furthermore, in all but the Altieri and Townsend (2011) audio-visual experiments, the stimuli were static.

Information fusion is primarily concerned with taking different sources of information and combining them in such a way that the least amount of information is lost. When the fused information is presented visually, for human use, the goal of preserving information is usually balanced with the desire to limit the amount of time or cognitive resources that an individual must dedicate to use the information. For example, displaying an image based on visible light and the same scene, but using heat, i.e. long-wave infrared, sensors next to each other is generally not considered a good solution, despite preserving all of the information available in each image. Our project focuses on combining information from visible light with information from infrared light sensors. Many algorithms are available for combining these sources of information into a single image, and assessing which algorithms are the best is an important aspect of image fusion. As mentioned in the introduction, many approaches to evaluating information are based on information theoretic or other metrics that are not directly based on human performance Toet et al. (2010a). The need to make the information usable to human operators has led to efforts to define metrics based directly on human performance. As assessing the fused information is, fundamentally, a question about how humans perceive and use multiple sources of information, SFT offers an ideal framework for studying this problem. First, the capacity coefficient offers a way to compare different combinations of information and different fusion algorithms using a common, cognitively plausible baseline for processing: The unlimited-capacity, independent, parallel system. Fusion algorithms that result in images that people perceive with super-capacity are the ideal, whereas algorithms that lead to limited capacity perception are to be avoided Repperger et al. (2009). Additionally, the SIC can be combined with the capacity coefficient to give insight into the processes that lead to variation in performance, i.e., which properties of the cognitive processes lead to super-capacity (e.g., coactive processing) or limited-capacity (e.g., serial processing) of fused information.

DISTRIBUTION A: Distribution approved for public release.

3 Advances in Data Analysis

This research program necessitate a number of advances in the statistical analysis of SFT constructs and of basic psychophysical measures. In the first subsection we discuss the development of Bayesian statistical tests for SFT. In the following subsection, we describe a hierarchical Bayesian tests for psychophysical thresholds and human information extraction efficiency.

3.1 Bayesian Systems Factorial Technology[1]

A number of approaches have been introduced for making inferences based on the SIC and MIC (see Houpt & Burns, 2016, for a review). The initial approach to testing the MIC values relied on using a factorial ANOVA design. ANOVA is an almost natural choice given that the factorial nature of an SFT study's manipulations. ANOVA is used to test the hypothesis on whether or not an observed MIC value significantly departs from zero value, which was identified as the null-hypothesis (Kirk, 2012). An alternative, nonparametric approach was to use bootstrapping (see Van Zandt, 2002, for details) to construct confidence intervals around observed MIC values. If zero is within the confidence intervals of the estimated MIC, a researcher would fail to reject null-hypothesis, otherwise the null is rejected and the sign of the MIC value determines whether the MIC shows overadditivity, or underadditivity (C.-T. Yang, Chang, & Wu, 2012; C.-T. Yang, Little, & Hsu, 2014, see e.g.,). An alternative, nonparametric test, based on a generalization of the Kolmogorov-Smirnov test, has also been proposed as an approach to analyzing the SIC shape, and hence whether the MIC is significantly different from 0 (Houpt & Townsend, 2010a). Houpt and Townsend (2010a) also compared standard ANOVA and nonparametric interaction tests for testing the null-hypothesis that $\mathrm{MIC} = 0$.

There are two main limitations of these existing approaches. The first limitation is related to the statistical inference and the diagnostic power of the SFT nonparametric methods. Although very useful at the initial stages of the development of the SFT technology, statistical inference based on null-hypothesis testing can be limiting. Using the ANOVA and bootstrapping approaches described above the null-hypothesis is exclusively linked to one mental architecture $\mathrm{MIC} = 0$, which is the signature of serial processing. A significant result would indicate that processing is not serial, but there is no way to reject parallel processing: If the result is non-significant, that does not imply there is evidence against $\mathrm{MIC} \neq 0$ (or for $\mathrm{MIC} = 0$).

Until recently the SFT approach has been focused on individual subject analysis, in addition to the statistical inference about the underlying cognitive operations. Indeed, the many SFT studies made a final results in the form of basic descriptive statistics, nominally classifying subjects based on their achievement. For example, a short-term memory study indicated individual differences

[1] Much of the content in this section is summarized or taken verbatim from Houpt, Heathcote, and Eidels (in press); Houpt, MacEachern, Peruggia, Townsend, and Van Zandt (2016) and Houpt and Fifić (2013).

DISTRIBUTION A: Distribution approved for public release.

across and within experimental conditions of different short-term memory manipulations. The major finding was that some subjects would switch from serial to parallel when the timing condition was changed (Townsend & Fifić, 2004). Although these results are very useful, the nominal categorization based on the statistical inferences using the null-hypothesis test, tells little about the population from which the subjects had been sampled.

The two limitations of statistical inference with SIC/MICs have been discussed, the first one being logically limited commitment to the null-hypothesis testing, and the second one being the lack of group level analysis. Both limitations can jeopardize the practical power of the SFT method, with possibility to systematically biasing inferences.

To address these limitations, we propose hierarchical Bayesian analysis. Hierarchical modeling allows for compromise between modeling individual differences and group level information (cf. Busemeyer & Diederich, 2010, Chapter 6). By employing a Bayesian approach, we can using priors to incorporate information about task constraints on a likelihood that some fundamental processing property are present. For example, when exhaustive processing is required by the task and accuracy is high, there is low prior probability that self-terminating processing was employed, hence MIC is less likely to be positive. A Bayesian approach can be used estimate posterior probabilities of each category of MIC (less than, greater than or equal to zero) rather than being limited to testing the null-hypothesis that MIC = 0. These MIC posterior probabilities can be estimated at both the individual level, indicating how likely each MIC category is for each subject, as well as the group level.

3.1.1 Parametric MIC

Our full model is depicted in Figure 3. The central component of the model is a linear model of the mean response time, much like an ANOVA (cf. Rouder, Morey, Speckman, & Province, 2012). We derived this linear model based on two principles. First, the MIC is the main variable of interest, so we needed it to be explicitly represented. This allows us to set priors on both its category and magnitude. Second, we ensured that the variability of the prior on the mean for each condition would not be different across the salience levels. There are a number of different possibilities for this matrix. For our purposes, we chose,

$$\begin{pmatrix} \text{MIC} \\ \Delta_2 \\ \Delta_1 \\ \mu \end{pmatrix} = \begin{pmatrix} 1 & -1 & -1 & 1 \\ -1/2 & 1/2 & -1/2 & 1/2 \\ -1/2 & -1/2 & 1/2 & 1/2 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix} \begin{pmatrix} \mu_{HH} \\ \mu_{HL} \\ \mu_{LH} \\ \mu_{LL} \end{pmatrix}.$$

Here, MIC is the mean interaction contrast, Δ_1 is the average increase in mean response time due to a change in salience on Channel 1 across salience levels on Channel 2 (and likewise for Δ_2), and μ is the grand mean response time.

Thus, if we set our priors on the MIC, Δ_i and grand mean, they can be translated into priors the mean RT at each salience level using the inverse of

DISTRIBUTION A: Distribution approved for public release.

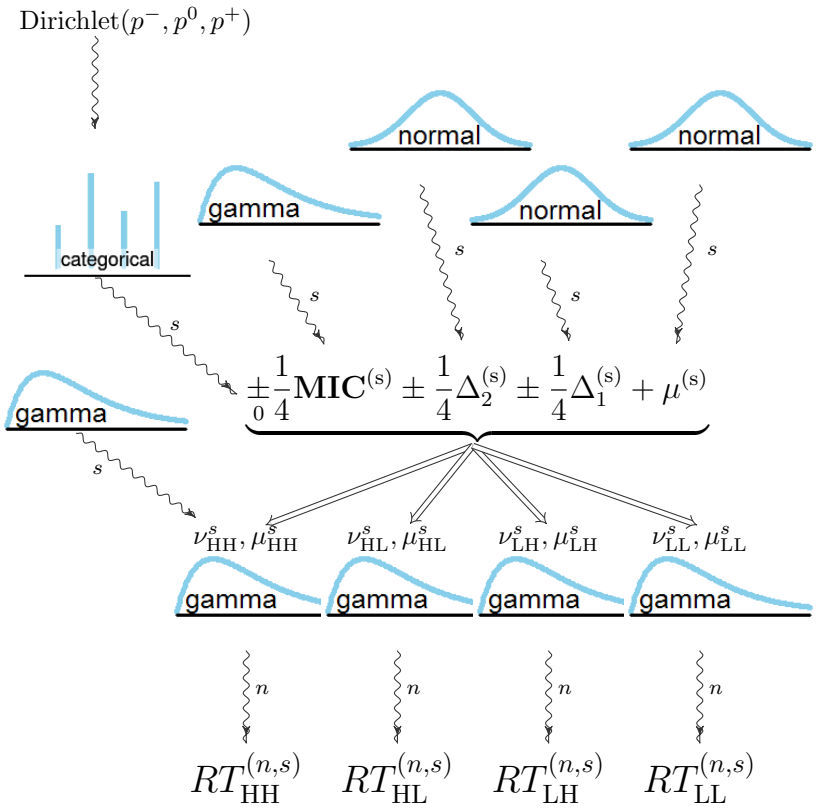

Figure 3: Diagram indicating the hierarchical structure of the Bayesian model of RT from which we deduce information about the MIC. Wiggling lines indicate a random relationship (e.g,. RTs are sampled from a gamma distribution), and double, straight arrows indicate a deterministic relationship (e.g., μ_{HH} is determined by the linear equation at the center of the diagram). Note that there is a separate prior for the rate parameter of each RT gamma distributions, although only one is depicted to reduce clutter. Thanks to John Kruschke (http://doingbayesiandataanalysis.blogspot.com/), Rasmus Bååth (http://www.sumsar.net/about.html) and Tinu Schneider (https://github.com/tinu-schneider/DBDA_hierach_diagram) for making this figure possible.

DISTRIBUTION A: Distribution approved for public release.

the mapping above,

$$\begin{pmatrix} \mu_{HH} \\ \mu_{HL} \\ \mu_{LH} \\ \mu_{LL} \end{pmatrix} = \begin{pmatrix} 1/4 & -1/2 & -1/2 & 1 \\ -1/4 & 1/2 & -1/2 & 1 \\ -1/4 & -1/2 & 1/2 & 1 \\ 1/4 & 1/2 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} \text{MIC} \\ \Delta_2 \\ \Delta_1 \\ \mu \end{pmatrix}.$$

To test the MIC, we set up a likelihood as a mixture model of three cases, one in which the MIC > 0, one with MIC < 0 and one with MIC $= 0$. Each subject's data had its own categorical distribution over the three cases with a Dirichlet prior over the case probabilities. Each of those Dirichlet priors were drawn from a single Dirichlet distribution representing the group level.

Assuming the RTs are on a millisecond scale, the prior on the magnitude of the MIC in the two cases for which it was non-zero, was a truncated Gaussian with mean 100 and standard deviation 50. Although a separate random variable was used for the MIC magnitude depending on whether it was for the positive case or negative case, all three three cases shared the same Δ and μ parameters. The priors on Δ_1 and Δ_2 each had a truncated Gaussian distribution with the same parameters (mean 100, standard deviation 50). For the grand mean μ, we use a truncated normal distribution with mean 400 and standard deviation 100.

In theory, the mean response time for particular subject in a condition could be negative under this model, e.g., when the effect of the salience manipulations has higher magnitude than the grand mean response time. Although this possibility should have no probability in the data, it is important to constrain the parameters of the prior distributions so that a negative mean response time is unlikely or impossible.

For the likelihood of the response times, we used a gamma distribution, which has the skewed shape commonly observer in RTs and only has support on positive values. The gamma distribution has two parameters, so to get from the mean response time as modeled above to the parameters of the gamma distribution, one additional parameter was required. In this case, we chose to use the rate parameter as the additional free parameter, then determine the shape from the product of the rate and mean RT. Like the Δ parameters, we used only a single rate parameter across the three MIC cases. For the analyses reported below, we chose improper flat priors over the positive real line for the rate parameters to allow flexibility in how the shape and rate traded off for a given mean RT.

To better understand how well this model can be used assess MIC category, and hence discriminate serial and parallel processing, we tested it on a series of simulated data. We varied the architecture and stopping rule for processing two sources of information, the factors of interest that determine the MIC category. Recall that selectively influenced serial models imply MIC $= 0$ regardless of stopping rule, parallel models with exhaustive stopping rules imply MIC < 0 and parallel models with self-terminating rules imply MIC > 0. In addition to the sign of the MIC, other factors can influence the magnitude of the MIC and precision with which it can be measured. One of the most important factors is

DISTRIBUTION A: Distribution approved for public release.

the effectiveness of the salience manipulation, i.e., how much faster each source of information is processed in a high salience condition relative to a low-salience condition. Additionally, the amount of data, particularly the number of response times collected from each subjects and the total number of subjects was varied.

Data were generated assuming either 10, 15, or 20 subjects. For each simulated subject, either 40, 50, 60, or 70 response times were simulated per condition (e.g., 70 in the HH condition, 70 in the HL condition, 70 in the LH condition, and 70 in the LL condition). Each response time was simulated by combining the subprocess durations (T_1, T_2) according the corresponding architecture and stopping rule:

Parallel, Exhaustive:	$\text{RT} = \max(T_1, T_2)$
Parallel, First-Terminating:	$\text{RT} = \min(T_1, T_2)$
Serial, Exhaustive:	$\text{RT} = T_1 + T_2$
Serial, First-Terminating:	$\text{RT} = T_1$ with probability $p = .5$ or $\text{RT} = T_2$ with probability $1 - p$

Subprocess durations were generated assuming the completion times were based on the first passage time of a Brownian motion process with diffusion coefficient 1, and hence followed an inverse Gaussian distribution,

$$f(t; \alpha, \nu) = \frac{\alpha}{\sqrt{2\pi t^3}} \exp\left[\frac{-(\alpha - \nu t)^2}{2t}\right].$$

The threshold activation for a response, α, was set to 30 and the diffusion coefficient, σ^2 for all simulations. The drift rate, ν, depended on the condition. To simulate a low salience trial for a subprocess, the drift rate was set to 0.1. For high salience trials, the drift rate was set to either $1.5, 2, 2.5$, or 3 times the low salience drift rate.

The Bayesian analyses were run using Stan (Stan Development Team, 2014a, 2014b) on a combination of MindModeling.org (Harris, 2008), the Oakley cluster at the Ohio Supercomputing Center (Ohio Supercomputer Center, 1987, 2012), and Microsoft's Azure service (*Microsoft Azure*, n.d.). Follow-up analysis were done using R statistical software (R Development Core Team, 2011) and the sft R package (Houpt, Blaha, McIntire, Havig, & Townsend, 2013). The Stan code is included as supplementary material.

A summary of the group level posterior and subject level posterior are shown in Figure 4 and Figure 5 respectively. Each row corresponds to a different model used to generate the data. The left column gives the mean posterior probability that the MIC is in the category predicted by the generating model (e.g., $MIC < 0$ for data generated from a parallel first-terminating model). The right column indicates the standard deviation of the posterior probability of that MIC category. In the subject level data, the values are averaged across the simulated subjects (i.e., the mean posterior probability is the average across subjects of their individual mean posterior probability; the standard deviation is the average across subjects of the standard deviation of the posterior probability that their MIC is in the given category).

DISTRIBUTION A: Distribution approved for public release.

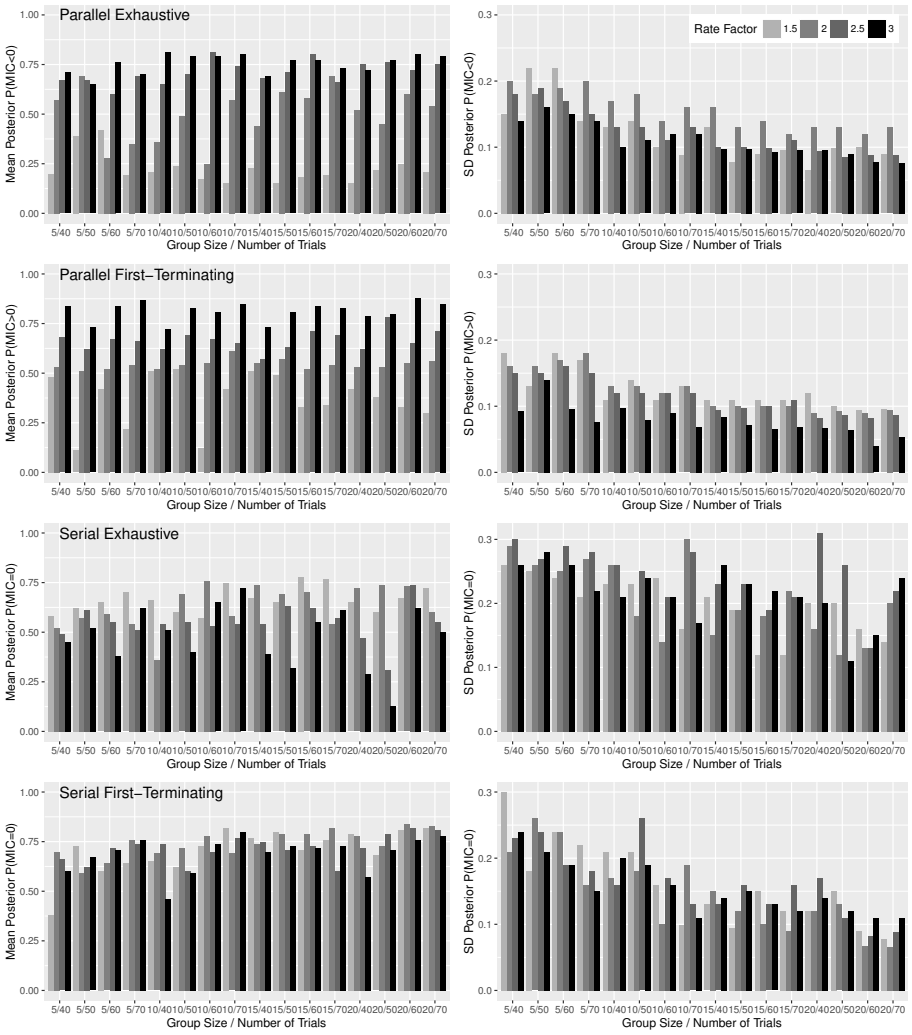

Figure 4: Simulation results for the group level probabilities. Each row corresponds to a model that was used to generate the data. The left column shows the mean posterior probability that the group MIC is in the category predicted by the model that generated the data. The right column shows the standard deviation of the that posterior probability. Within each panel, bars are grouped by the number of trials per subject, then by the number of subjects per group. The rate factor, representing the strength of the salience manipulation, is indicated by the shade of the bars.

DISTRIBUTION A: Distribution approved for public release.

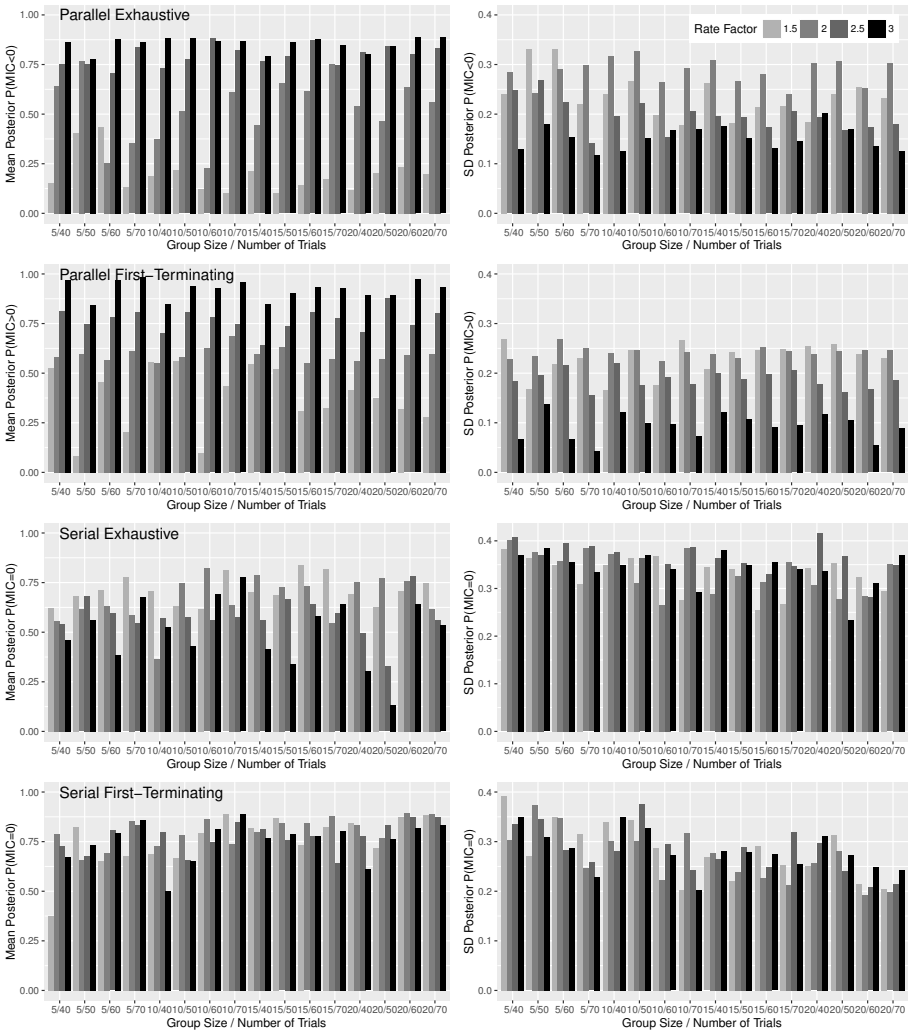

Figure 5: Simulation results for subject level probabilities. As in the previous figure, row corresponds to generating model. In this figure, the left column shows the posterior probability that the subject MIC is in the category predicted by the model that generated the data, averaged across subjects. The right column shows the standard deviation of the that posterior probability averaged across subjects. Within each panel, bars are grouped by the number of trials per subject, then by the number of subjects per group. The rate factor, representing the strength of the salience manipulation, is indicated by the shade of the bars.

DISTRIBUTION A: Distribution approved for public release.

The only factor that had a clear effect on the posterior probability over MIC category, for both the group and individual level, was the strength of the salience manipulations (indicated by line darkness in Figures 4 and 5). At the lowest manipulations strength, the most likely MIC is 0 for all of the models, regardless of the number of subjects or the number of trials per distribution. The posterior probability correct of positive and negative MICs increases essentially linearly with an increase in salience for the parallel–exhaustive data and parallel–first-terminating data respectively. In the serial, first-terminating data, the posterior probability stays essentially flat between 0.6 and 0.8 for the range of salience. Interestingly, there seems to be a negative trend in the serial-exhaustive data, particularly with only 50 trials per distribution.

The standard deviation of the category probability was also affected by the salience strength. In the parallel model data, the lowest rate factor resulted in lower standard deviations, reflecting more certainty in the posterior that the MIC was zero. For the larger rate factors in the parallel data, the lower standard deviation was again smaller, but in this case reflecting more certainty that the MIC was negative or positive for the exhaustive and first-terminating data respectively.

In addition to the rate factor, the number of subjects affected the group level and the number of trials per subject affected the subject level. More subjects led to lower standard deviations at the group level, and lower standard deviations at the subject level, although the effect was more prominent at the group level. More trials per subject led to lower standard deviations at the subject level, but had little affect at the group level.

In general, we find these results quite promising. Most experiments relying on SICs use at 100 or more trials per distribution and approximately 10 subjects (e.g., C.-T. Yang et al., 2011). Our results indicate that, as long as the salience manipulation is sufficient, this is enough data for drawing both group and individual level inferences. The results regarding the rate factor indicate an important cautionary note as well. If the salience manipulation is not strong enough, data from any of the four generating models will be classified as having a zero MIC. Hence it is important to aim for strong salience manipulations in designing experiments to be analyzed with this (or any other SIC) analysis. Based on the impact of the rate factor, when the salience is strong, the model should do well.

One of the standard data sets for testing SFT statistical analyses is the dot detection data reported in Eidels, Townsend, Hughes, and Perry (2015, Study I) and available in the sft R package (Houpt, Townsend, & Donkin, 2014; R Development Core Team, 2011). In this study, one or two small, low-contrast dots were shown on a uniform background either above the mid-line of the display, below the mid-line, or both. Each dot could be displayed at a slightly higher contrast (high salience) or lower contrast (low salience). There were three factors manipulated within subjects: dot presence (present, absent); dot salience (high, low); and task instructions (OR and AND). The task instructions were held constant within a day. For example, on one day participants were asked to respond affirmatively if either dots was shown and negatively otherwise (OR

DISTRIBUTION A: Distribution approved for public release.

rule). On another day, participants were asked to respond affirmatively only if both dots were shown and negatively otherwise (AND rule).

The simple detection study allows for the model assessment by inspecting the observed MIC values. If the participants were processing the visual stimuli in parallel, we would expect a positive MIC in the OR condition and a negative MIC in the AND condition. It is also possible that despite the "OR" instructions, the participants used an exhaustive stopping rule in that condition, in which case we would expect a negative MIC in both conditions. In the AND condition, the participants would have low accuracy if they used a first-terminating stopping rule, which was not the case. However, if participants were using a coactive strategy, then a positive MIC would be indicated in the AND condition. If a participant used a serial strategy, either exhaustive or first-terminating, the resulting MIC would be 0. For estimating the MIC, there were 200 trials for each condition of interest (HH, HL, LH, and LL) for each instruction type. The data set provided results that are consistent across subjects, and clearly identifiable using the SFT approach. As such the data set provides a valuable validation tool for the new analysis.

In our initial application of the new hierarchical analysis to the Eidels et al. (2015) data, we separately analyzed the AND condition and the OR condition. As in the simulations section, we ran four chains using $10,000$ warm-up samples and $20,000$ additional iterations per chain. All chains were visually assessed for mixing and Gelman-Rubin $\hat{R}$ values were less than 1.01 for all parameters.

Results of the first analysis are reported in Table 1 and are consistent with previous analyses based on non-Bayesian methods (Houpt et al., 2016; Houpt & Townsend, 2010a). For the AND task, the posterior probabilities strongly favored the negative MIC at the group level and for each of the individuals. Similarly, for the OR task, positive MICs had the highest probability at the group level and for each of the individuals. Two participants, S2 and S4, had relatively lower probabilities of positive MICs in the OR task, with posterior odds ratios of 2.8 and 10 respectively for positive over zero MICs. On the whole, there is strong evidence against serial processing (which implies MIC $= 0$). Further, there is even stronger evidence against coactive processing in the AND task ($MIC > 0$) or exhaustive processing in the OR task (MIC < 0).

Given that the model indicated the same MIC category across participants, one may wonder whether the hierarchical model is biased toward assuming a single MIC category for all participants. While a bias toward homogeneity could be intentionally built into the model by using a group level prior with most of the probability mass focused on a particular MIC category, the prior we used was meant to allow variability across subjects. To explore the possibility that the model is biased toward homogeneity, we recoded the Eidels et al. (2015) data so that each participant–instruction combination was treated as a separate member of a single group. I.e., the data from Subject 1 in the OR condition was recoded as S1-OR while the data from him in the AND condition was recoded as S1-AND and both, and likewise for the other 8 participants.[2] We ran four chains

[2]Although we could have built structure into the model relating a subject's performance

DISTRIBUTION A: Distribution approved for public release.

	AND Task			OR Task		
	+	0	−	+	0	−
Group	0.14	0.06	0.80	0.73	0.11	0.17
S1	0.04	0.00	0.95	0.93	0.01	0.06
S2	0.04	0.06	0.90	0.53	0.38	0.09
S3	0.06	0.02	0.92	0.94	0.01	0.05
S4	0.02	0.02	0.95	0.63	0.27	0.10
S5	0.02	0.02	0.97	0.94	0.04	0.02
S6	0.06	0.11	0.83	0.95	0.03	0.02
S7	0.03	0.04	0.93	0.78	0.15	0.07
S8	0.05	0.08	0.87	0.92	0.05	0.03
S9	0.06	0.01	0.94	0.93	0.03	0.04

Table 1: Mean posterior probabilities of MIC category when AND and OR conditions were analyzed separately.

using $10,000$ warm-up samples and $20,000$ additional iterations per chain. All chains were visually assessed for mixing and Gelman-Rubin $\hat{R}$ values were less than 1.01 for all parameters.

The posterior probabilities in Table 2 indicate very little probability of a zero MIC, but roughly equal probabilities of positive and negative MICs at the group level. This is noteworthy for two reasons: First, it demonstrates that the model does not inherently predict homogeneity. Second, it illustrates the advantage of using a categorical prior for the sign of the MIC because the positive and negative individual MICs were not averaged (which would give a group MIC near zero). Despite the fact that the posterior probabilities indicate heterogeneity, there was still some shrinkage in the individual posterior probabilities: For the AND data, the probability of a negative MIC was slightly smaller and slightly larger for positive MICs while the opposite was true for the OR data. The probability of a zero MIC stayed was roughly the same for the AND data as in Table 1. The probability of a zero MIC in the OR data decreased some, particularly for those participants for who had slightly higher posterior probabilities of a zero MIC on the OR condition in Table 1. It is clear that this model does not impose homogeneity on the individuals.

On the whole, these results are quite promising. The model was able to estimate a reasonable group level and individual level posterior distribution. These results provide converging evidence with the previously reported analyses of these data, which had shown parallel processing for all participants and the appropriate stopping rule application for the specific stopping rule task instruction condition. The additional benefit of the new Bayesian hierarchical approach is that it provides not only the individual level information, but also the group level information.

The survivor and mean interaction contrasts are among the most powerful

across the instructions, we chose to treat the RTs for a given subject with a given instruction as conditionally independent given the group MIC value.

DISTRIBUTION A: Distribution approved for public release.

	+	0	−
Group	0.46	0.06	0.48
S1-AND	0.04	0.00	0.95
S2-AND	0.09	0.05	0.86
S3-AND	0.09	0.02	0.90
S4-AND	0.06	0.02	0.93
S5-AND	0.04	0.01	0.94
S6-AND	0.15	0.11	0.74
S7-AND	0.07	0.04	0.89
S8-AND	0.12	0.07	0.81
S9-AND	0.07	0.01	0.92
S1-OR	0.93	0.01	0.06
S2-OR	0.56	0.28	0.17
S3-OR	0.94	0.01	0.05
S4-OR	0.58	0.17	0.25
S5-OR	0.93	0.02	0.05
S6-OR	0.93	0.02	0.05
S7-OR	0.77	0.08	0.15
S8-OR	0.91	0.02	0.07
S9-OR	0.93	0.01	0.06

Table 2: Mean posterior probabilities of MIC category when AND and OR conditions were analyzed as a samples from the same group distribution. Si-AND indicates data from the AND instructions while Si-OR indicates data from the OR instructions. The model did not encode the relationship between AND and OR data from the same subject.

DISTRIBUTION A: Distribution approved for public release.

diagnostic methods for discerning whether people using information in parallel or in series because they avoid the model mimicking dilemma that plagues other methods. However, the interaction contrast approach complicates the statistical analysis so methods for statistical inference have been relatively lacking until recently. Houpt and Townsend (2010a) proposed a null-hypothesis test for the SIC and compared ANOVAs and adjusted-rank-transform tests for the MIC.

In this project we addressed one of the major outstanding issues in the statistical analysis of MICs, the lack of an approach to make group level inferences. We demonstrated the efficacy of a hierarchical Bayesian model of the MIC for making both individual level and group level inferences with a relatively small number of trials and subjects, using both a simulation study and an application to a standard dataset. Performance of the analysis on the simulated data improves with having more subjects, trials, and increased efficacy of the salience manipulation. Nonetheless, with just 50 trials per condition, inferences based on the model's posterior probability of of the MIC associated with the data generating process led to quite satisfactory results.

Both the SIC and MIC measures are frequently used as an individual subject assessment to indicate qualitative differences in cognitive operations in a sample of subjects. An obstacle in assessment of individual human subjects' cognitive operations is the requirement for a large number of trials per subject. For example, Houpt and Townsend (2010a) demonstrated their statistical analysis with 200 trials per distribution, which when trials are balanced appropriately (cf. Houpt & Townsend, 2012; Mordkoff & Yantis, 1991a) can mean 3200 trials per participant. While this samples size would not cause a psychophysicist to balk, many interesting populations, such as clinical groups, experts, and some age groups, are available only for a limited time, and thus permit only a smaller set of observations per individual. Although it has less diagnostic power then SIC, the MIC can rely on a small data sets, making it a more practical measure for cases in which only limited numbers of trials are available per subject.

One unexpected finding was that with increased salience manipulation efficacy but a limited number of trials, MIC category recovery performance weakened for the data generated from a serial exhaustive process. As the stimulus salience effect increased, the posterior probability of a zero MIC decreased. The extent to which this is a property of the particular assumptions we have made, either in generating the data or the model itself, or if it is an outcome specific to this sample dataset, will be an interesting topic of further investigation.

In addition to the simulated data, the model performed well on the SFT data that is commonly used to assess SIC and MIC statistics from Eidels et al. (2015). The Bayesian hierarchical MIC model exhibited strong convergence to the conclusions drawn from SIC level analysis in other papers (Houpt et al., in press, 2016; Houpt & Townsend, 2012). Perhaps the most challenging test of the model was its application to heterogeneous experimental conditions in which the subjects were using different processes. In the Eidels et al. (2015) study, two experimental conditions were imposed by the instructions. In the OR condition, subjects could use a self-terminating stopping rule, while in the AND condition they have to use an exhaustive stopping rule. To test whether the model is

DISTRIBUTION A: Distribution approved for public release.

able to detect variation across subjects, the data in each condition were treated as coming from the same group, thus having heterogeneous subject properties. When the hierarchical Bayesian MIC model was applied to the data in this format, the analysis appropriately identified the expected MIC category at the subject level and indicated approximately 50% posterior probability for each of the positive and negative MIC categories at the group level. This demonstrated that the Bayesian MIC model can identify individual subjects' differences within a group data set, and will not always indicate that all subjects use the same cognitive operations.

Our approach to exploring the individual and group level MIC analysis using the hierarchical Bayesian MIC model is similar in many ways to the method proposed by (Thiele & Rouder, 2016).[3] The overarching goals of both approaches are the same: 1) To better quantify the evidence for either serial or parallel processing at the group level 2) Rein in the bias toward heterogeneity that results from analyzing subjects as unrelated. Similarly, the structure of the models are quite similar, with a linear model predicting the mean processing time across distributions within a subject. There are three main differences between the two models. First, Thiele and Rouder (2016) use a normal distribution as their model of the response times where as we use a gamma distribution. They report choosing the normal distribution for two reasons, computational tractability and the ease with which the sign of the MIC can be constrained relative to non-normal distributions. From our perspective, the computational power of Stan and Hamiltonian Monte-Carlo methods means that we can use a more realistic distribution for response times and still obtain results from the analysis in a reasonable time frame. Furthermore, our categorical approach using the Dirichlet prior allows us to model the sign of the MIC without additional difficulty in implementation. The second difference between the approaches is the means by which conclusions are drawn. The Thiele and Rouder (2016) approach focuses on pairwise Bayes factor comparisons between models with the MIC either constrained to be positive, negative or zero. We use the categorical distribution to represent whether the MIC is positive, negative or zero. On the surface, this amounts to only a trivial difference as the Bayes factor can easily be calculated from the categorical priors and *vice versa*. The advantage of our approach is that the categorical distributions afford a hierarchical representation of the MIC category. This allows us to directly examine both the posterior probability that the MIC is a certain category at the group level and at the subject level. Posterior inferences regarding different MIC categories at the individual level possible in principle with Thiele and Rouder (2016) model in which each individual's MIC category is independently sampled from a normal prior distribution. One potential challenge for their approach is that differences across subjects are treated as ratio scale rather than categorical, hence a clear subset of participants with positive MIC and another subset with negative MIC would be treated as uncertain evidence for an average zero MIC.

[3]Both research groups independently developed research approaches to extending the SFT MIC tests using the hierarchical Bayesian model, and discovered each others work through presentations of their early results at the annual meeting of the Psychonomics Society.

DISTRIBUTION A: Distribution approved for public release.

3.1.2 Parametric SIC

Parametric models of information processing assume specific distributions of the time taken by the system to complete processing of some input and decide between response alternatives. The processing systems, or models, discussed in this paper combine input from multiple channels so the distributions can characterize either the completion time of each individual channel, or the total completion time of the system. For illustration, suppose there are two processing channels 1 and 2 that take time T_1 and T_2, respectively, to complete processing. These completion times for the individual channels are typically not observed. The observed quantity is the *total* completion time of the system, which depends on not only T_1 and T_2 but also the architecture and stopping rule. For example, a parallel minimum-time (Parallel-OR) model assumes that channels 1 and 2 are processed simultaneously and the system can stop as soon as the faster channel completes. Thus, the observable total completion time of such a system, when presented with input signals to both channels 1 and 2, can be mathematically expressed by $\min(T_1,\ T_2)$. In contrast, a parallel exhaustive (Parallel-AND) model assumes the system must await the slower of the two processes, so total completion time is $\max(T_1,\ T_2)$. A serial-minimum-time (Serial-OR) model assumes that channels are processed one after the other, but only one of them needs to be processed so total completion is equivalent to one channel's completion time, T_1 or T_2, depending on which channel is processed first (i.e., Channel 1 with probability p, or Channel 2 with probability $1 - p$). Finally, a serial-exhaustive model (Serial-AND) requires that channels are processed one after the other and that both must be completed for the system to finish, so total completion time is the sum of the completion-times of the individual channels, $T_1 + T_2$. For convenience we summarize the models below:

Parallel-OR	$T_{12} = \min(T_1, T_2)$
Parallel-AND	$T_{12} = \max(T_1, T_2)$
Serial-OR	$T_{12} = T_1$; with probability p, or
	$T_{12} = T_2$; with probability $1 - p$
Serial-AND	$T_{12} = T_1 + T_2$

These equations describe the completion time of four of the models of interest. They can be used to mathematically characterize the processing models with multiple processing channels based on the completion time *distributions* of each channel. Those distributions are combined in accordance with the model's architecture and stopping rule to get a model of the total completion time distribution, corresponding to the observed RT distribution (excluding the time for response production).

The formulas for distributions of completion times given the channel completion times are relatively straightforward in the two-channel case, but require a few definitions. Let $f(t)$ be the completion time density function, indicating the probability density that a process has finished at time t. Let $F(t)$ be the cumulative distribution function of processing time, indicating the probability that the process of interest has finished at time t or before. Finally, as

DISTRIBUTION A: Distribution approved for public release.

already discussed, $S(t)$ is the survivor function, indicating the probability that the process had not finished at time t.

A Parallel-OR model implies the total completion time density is the density of channel 1, f_1(t), times the survivor function of channel 2, plus the density of channel 2 times the survivor function of channel 1. A Parallel-AND model has the same form, with the survivor function replaced by the cumulative distribution function. The density of the Serial-OR completion time is a mixture of the channel completion time densities, with the mixing proportions (p and $1-p$) equal to the probabilities that either channel goes first. Finally, the Serial-AND completion time density is the convolution ("*") of the two channel completion times (Townsend & Ashby, 1983a).

Let the subscripts 1 and 2 indicate the completion time of channels 1 and 2 respectively, and let the subscript 12 indicate the total completion time. Given $f_{12}(t)$ is the density of the total completion time, these models can be summarized as follows.

Parallel-OR	$f_{12}(t) = f_1(t)[1 - F_2(t)] + f_2(t)[1 - F_1(t)]$
Parallel-AND	$f_{12}(t) = f_1(t)F_2(t) + f_2(t)F_1(t)$
Serial-OR	$f_{12}(t) = pf_1(t) + (1-p)f_2(t)$
Serial-AND	$f_{12}(t) = f_1(t) * f_2(t)$

Note that the Serial-OR model has an extra parameter, p, that determines the probability that either of the two channels is processed first. For the purposes of this paper, we left this as a free parameter with a uniform prior between 0 and 1.

Characterization of the fifth model of interest, the coactive model, depends on the underlying process that generates its completion time distribution. Often RTs are modeled as the amount of time until a stochastic process first reaches a particular threshold value. The stochastic process represents the accumulation of evidence for a particular response, and the threshold represents the amount of evidence required to make that response. Within this framework, a coactive model of RTs is based on the first passage time of the *sum* of the information accumulation processes on each channel to a single threshold. The most mathematically well developed coactive models are those based on a Poisson process (Schwarz, 1989; Townsend & Nozawa, 1995a) or on a Brownian motion process with drift (Houpt & Townsend, 2011; Schwarz, 1994).

Our test of the coactive model assumed that information accumulation is a Brownian motion processes with drift (ν) and threshold value (α). With these assumptions, the channel completion times have an inverse Gaussian distribution, with the following density and distribution functions.

$$f(t;\nu,\alpha) = \sqrt{\frac{\alpha^2}{2\pi t^3}} \exp\left[\frac{-(t\nu-\alpha)^2}{2t}\right]$$

$$F(t;\nu,\alpha) = \Phi\left[\sqrt{\frac{\alpha^2}{t}}\left(\frac{t\nu}{\alpha}-1\right)\right] + \exp\left[2\alpha\nu\right]\Phi\left[-\sqrt{\frac{\alpha^2}{t}}\left(\frac{t\nu}{\alpha}+1\right)\right]$$

DISTRIBUTION A: Distribution approved for public release.

Although a Brownian motion process could vary based on both a drift rate and a diffusion coefficient (i.e., the moment-to-moment standard deviation), we fixed the diffusion coefficient to 1 in our simulations. The diffusion coefficient can trade off with the drift rate and threshold, so one of them must be fixed to make the completion time distribution identifiable. We also assume that the drift rate at a given salience level is the same for each channel. Although this assumption will sometimes be incorrect, it reduces the complexity of the model, and hence the time needed to estimate posteriors. In many applications, such as the dot detection task that we use to illustrate applying tests to real data, this assumption is reasonable.

To fully specify the coactive model we use gamma prior distributions over the rate and threshold parameters. To constrain the model such that the high salience rate is larger than the low salience rate, we first sample the low salience drift rate, then add an additional, positive random variable (η) to get the high drift rate. Here we use the rate/threshold parameterization of the inverse Gaussian so that $\mathcal{IG}(\nu, \alpha)$ refers to the first passage time (i.e., the first time at which the threshold is reached) distribution for a Brownian motion process with drift rate ν to reach α. Note that this model is a first-passage time model and hence there is only one response-generating threshold per process. Figure 6 demonstrates the IG model of a Parallel-AND process and a Serial-AND process.

$$
\begin{aligned}
T_{i;H} &\sim \mathcal{IG}\left(\nu_H, \alpha\right) & \eta &\sim \text{Exponential}(100) \\
T_{i;L} &\sim \mathcal{IG}\left(\nu_L, \alpha\right) & \nu_L &\sim \Gamma(4, 0.1) \\
\alpha &\sim \Gamma(4, 0.1) & \nu_H &= \nu_L + \eta
\end{aligned}
$$

A growing number of software packages are available for estimating posterior distributions of a wide range of models, including those with a linear or nonlinear hierarchical structure. We implemented these models using STAN (Stan Development Team, 2014a), which is based on Hamiltonian Monte Carlo sampling (Duane, Kennedy, Pendleton, & Roweth, 1987).

3.1.3 Semiparametric SIC

To begin, recall that, for a vector of parameters $\boldsymbol{\theta}$, the likelihood of data $\boldsymbol{T} = \{t_1, t_2, \ldots, t_n\}$ is given by

$$
L(\boldsymbol{\theta} \mid \boldsymbol{T}) = \prod_{k=1}^{n} f(t_k \mid \boldsymbol{\theta}),
$$

when $\boldsymbol{T}$ is assumed to be an independent and identically distributed sample drawn from a distribution with pdf f. As we defined above, the hazard function for the variable T_k is

$$
h(t \mid \boldsymbol{\theta}) = f(t \mid \boldsymbol{\theta}) \left(S(t \mid \boldsymbol{\theta})\right)^{-1},
$$

DISTRIBUTION A: Distribution approved for public release.

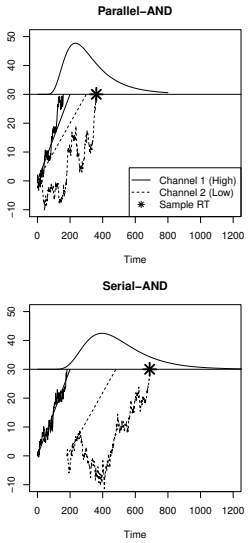

Figure 6: Illustration of the inverse Gaussian parametric model. The top panel depicts a Parallel-AND process, in which both processes accumulate simultaneously and a response is executed when the later of the two processes reaches its threshold. The bottom panel illustrates a Serial-AND process, in which the second process does not start to accumulate until the first process has reached its threshold and a response is executed when the second process reaches its threshold. In both panels the diagonal lines originating at 0 indicate the average accumulation while the jagged lines depict a sample accumulation process. The distribution of response times is indicated above the threshold. For clarity, the threshold for both channels is depicted as the same although the model does not have that constraint.

DISTRIBUTION A: Distribution approved for public release.

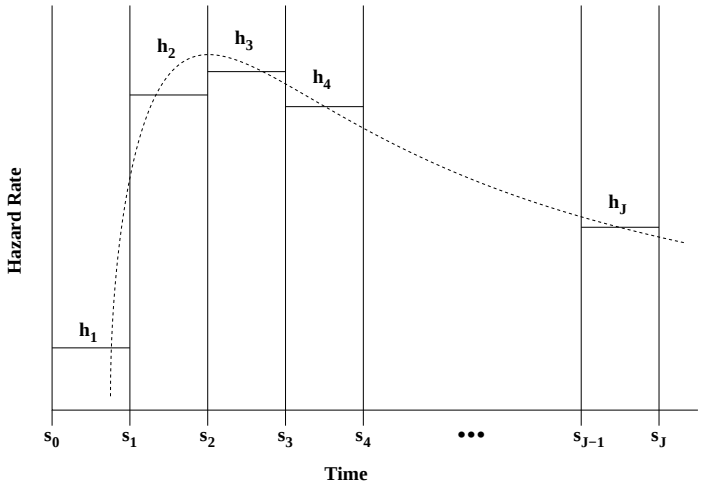

Figure 7: The piecewise exponential approximation of a continuous hazard function (dotted line). The approximation is represented by the horizontal line segments with heights h_j between the vertical boundaries at s_j, $j = 0, \ldots, J$.

where we will assume implicitly that the support of T_k is the positive real line, and that f takes on the value 0 for all non-positive values of t. Noting that

$$f(t \mid \boldsymbol{\theta}) = h(t \mid \boldsymbol{\theta}) S(t \mid \boldsymbol{\theta}),$$

we can rewrite the likelihood of the data $\boldsymbol{T}$ as

$$L(\boldsymbol{\theta} \mid \boldsymbol{T}) = \prod_{k=1}^{n} h(t_k \mid \boldsymbol{\theta}) S(t_k \mid \boldsymbol{\theta}).$$

We want to estimate values of the function $h(t \mid \boldsymbol{\theta})$ at a finite number J of points over the observed range of T_k. To do this we will use a semiparametric approach in which we assume that the behavior of the hazard function at these points can be well described by a piecewise exponential model (Ibrahim, Chen, & Sinha, 2005), depicted in Figure 3.1.3. We first place $J+1$ points $\{s_0, s_1, \ldots, s_J\}$ along the support of T_k, where $s_0 = 0$ and $s_J > \max_k\{T_k\}$. The hazard function $h^*(t)$ of the piecewise exponential model is piecewise constant, i.e.,

$$h^*(t) = h_j \text{ for } s_{j-1} < t \leq s_j,\ j = 1, \ldots, J.$$

Given the values of the hazard function, $\boldsymbol{h} = \{h_1, h_2, \ldots, h_J\}$, the piecewise exponential pdf is

$$f(t \mid \boldsymbol{h}) = \begin{cases} h_j \exp\left\{-h_j(t - s_{j-1}) - \sum_{m=1}^{j-1} h_m(s_m - s_{m-1})\right\} & \text{for } s_{j-1} < t \leq s_j,\ j = 1, \ldots, J \\ 0 & \text{for } t \notin [0, s_J). \end{cases} \tag{6}$$

DISTRIBUTION A: Distribution approved for public release.

Note that the parameters of the model are the hazard rates $\{h_1, h_2, \ldots, h_J\}$ that define the exponential rates within each bin. The likelihood can then be rewritten as

$$\begin{aligned} L(\boldsymbol{h} \mid \boldsymbol{T}) &= \prod_{k=1}^{n} f(t_k \mid \boldsymbol{h}) \\ &= \prod_{k=1}^{n} \prod_{j=1}^{J} \left[h_j \exp\left\{ -h_j(t_k - s_{j-1}) - \sum_{m=1}^{j-1} h_m(s_m - s_{m-1}) \right\} \right]^{I(s_{j-1} < t_k \leq s_j)} \end{aligned} \tag{7}$$

Over conditions (c) and subjects (i), our goal is to estimate the posteriors of $h_{ic,j}$, $j = 1, \ldots, J$, and thus estimate the posteriors of the hazard functions of a hierarchical piecewise exponential model.

To fully specify the model, note that an RT T_{ick} from Subject i in Condition c on Trial k is distributed as a piecewise exponential with hazard rates $\boldsymbol{h}_{ic} = \{h_{ic,1}, h_{ic,2}, h_{ic,3}, \ldots, h_{ic,J}\}$. Given $h_{ic,1}$, the hazard rates $h_{ic,j}$ for $j = 2, \ldots, J$ are defined as a stationary autoregressive process of order 1 (AR(1)) on the log scale, so

$$\ln h_{ic,j} = \mu_{ic} + \phi_{ic} \left(\ln h_{ic,j-1} - \mu_{ic} \right) + \epsilon_{icj}. \tag{8}$$

The parameter $\mu_{ic} \in \mathbb{R}$ is the mean of the AR(1) process, $\phi_{ic} \in \mathbb{R}$ is the autoregressive coefficient, and the ϵ_{icj}s are normal innovations with mean 0 and variance σ_h^2.

We modeled the autoregressive parameter ϕ_{ic} as a rescaled inverse logit transformation of a parameter α_{ic} where

$$\alpha_{ic} \sim \mathcal{N}(0, \sigma_\alpha^2).$$

Specifically,

$$\phi_{ic} = 2/\left(1 + \exp(-\alpha_{ic})\right) - 1,$$

which restricts the autoregressive parameter ϕ_{ic} to the interval $(-1, 1)$, which in turn enforces stationarity of the AR(1) process. We selected normal priors for $\ln \boldsymbol{h}_{ic}$, such that

$$\ln h_{ic,j} \mid \ln h_{ic,j-1} \sim \mathcal{N}\left[\mu_{ic} + \phi_{ic}(\ln h_{j-1} - \mu_{ic}), \sigma_h^2 \right]$$

and

$$\ln h_{ic,1} \sim \mathcal{N}\left[\mu_{ic}, \sigma_h^2/(1 - \phi_{ic}^2) \right],$$

the latter following from the stationarity of the AR(1) process.

To construct the full hierarchical model over subjects and conditions, we gave μ_{ic} and α_{ic} normal priors so

$$\mu_{ic} \sim \mathcal{N}(\mu_{\mu,i}, \sigma_\mu^2) \text{ and } \alpha_{ic} \sim \mathcal{N}(\mu_{\alpha,i}, \sigma_\alpha^2),$$

where

$$\mu_{\mu,i} \sim \mathcal{N}(\mu_0, \sigma_{0,\mu}^2) \text{ and } \mu_{\alpha,i} \sim \mathcal{N}(\alpha_0, \sigma_{0,\alpha}^2).$$

DISTRIBUTION A: Distribution approved for public release.

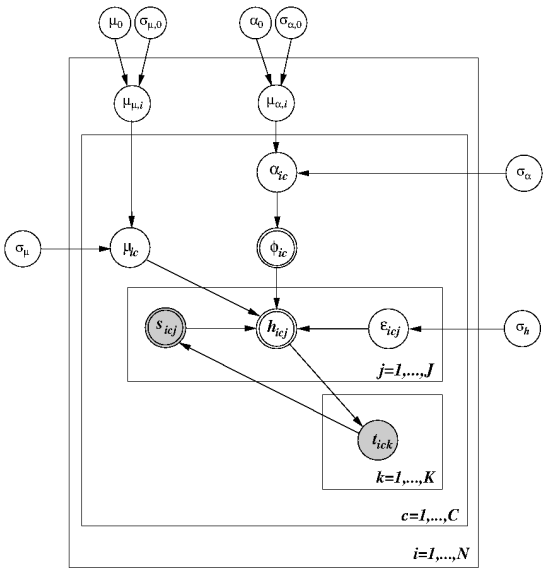

Figure 8: The graphical model of the semiparametric hazard function estimate. The index j runs over the J bins in which the hazard rates are estimated, k represents trials, c represents experimental conditions and i represents individuals. Continuous parameters to be estimated are shown in single circles, continuous computed values in double circles, and observable variables in shaded circles.

To complete the model, we gave the hyperparameters normal and gamma priors[4], so

$$\mu_0, \alpha_0 \sim \mathcal{N}(0, 1),$$
$$\sigma_\mu^2, \sigma_\alpha^2, \sigma_h^2 \sim \Gamma(1, 1),$$

and

$$\sigma_{0,\mu}^2, \sigma_{0,\alpha}^2 \sim \Gamma(1, 1).$$

With this model structure in place, it is important to note that the prior mean for $\boldsymbol{h}_{ic}$ is $\mathbf{1}$, or that the prior mean shape of the hazard is that of the exponential distribution with rate parameter 1. Figure 8 shows a graphical representation of this model.

It is an open question as to the best way to select the bin boundaries $s_{ic,j}$. There are three possibilities that we considered: 1) using fixed, equidistant points between $s_{ic,0} = 0$ and some reasonable choice for $s_{ic,J} > \max_k T_{ick}$; 2) using the observed quantiles of the samples $\boldsymbol{T}_{ic}$; and 3) incorporating the values

[4]Because the hazard rates generally varied over a narrow range, depending on the unit of measurement for T_{ick}, priors with relatively high precision resulted in priors for $\boldsymbol{h}$ with low precision.

DISTRIBUTION A: Distribution approved for public release.

of $\boldsymbol{s}_{ic}$ into the structure of the model and estimating their posteriors along with the posteriors of $\boldsymbol{h}_{ic}$.

Using fixed, equidistant points has much to recommend it. Unfortunately, for distributions with heavy tails, often observed for RT data, the value of $s_{ic,J} > \max_k T_{ick}$ is likely to be distant from most of the values in the sample. Slicing the range from 0 to $s_{ic,J}$ into J bins of equal width results in a potentially large number of bins in the tail of the distribution within which no observations were sampled. There seemed to be no practical way around this problem, even for small numbers of subjects in small numbers of conditions.

We explored the inclusion of $\boldsymbol{s}_{ic}$ among the model parameters by defining a new parameter vector $\boldsymbol{p}_{ic} \in (0,1)^J$ with a Dirichlet prior, and setting $s_{ic,q} = (1+\delta)\max_k T_{ick} \sum_{m=1}^{q} p_{ic,m}$ for small values of δ. The increase in the dimensionality of the model resulted in unstable estimates which led to an inability to fit hierarchical structures in a practical way.

The failure of these two methods left us with the quantile solution. For J hazard rates, we used the J/N quantiles of the sample which assured that there were N/J observations within each bin. This approach also has drawbacks, namely that the estimates of $\boldsymbol{h}_{ic}$ reflect in part the possible variance in the bin boundaries $\boldsymbol{s}_{ic}$, and that we are using the data twice, once to estimate the bins and then to estimate the hazard rates. Despite these drawbacks, the procedure worked well, and permitted us to obtain accurate posterior estimates of $\boldsymbol{h}_{ic}$ in reasonable amounts of time.

3.1.4 Nonparametric SIC

In this section, we develop a nonparametric Bayesian test of architecture and stopping rule. Rather than modeling the underlying process by making specific assumptions about channel time distributions, as we do in the test developed in the next section, here we directly model the uncertainty in the estimates of the survivor function in each condition, and hence the uncertainty in the estimate of the SIC. For Bayesian analysis, we need a prior over distributions. In order to not inadvertently constraint the generality of our characterization of a particular architecture and stopping rule we use the Dirichlet process prior (Ferguson, 1973).

To describe the Dirichelt process formally, it helps to begin with the Dirichlet distribution. The Dirichlet distribution is a distribution over the probabilities associated with a fixed number of bins. For example, if there were k categorical outcomes possible, then the Dirichlet distribution would describe the relative likelihood of an assignment of probability across those k bins. The parameter of the Dirchlet distribution, the vector β, effectively assigns prior weight to distributions that have bin likelihoods that are similar in relative magnitudes to the relative magnitudes of the elements of β. With k bins, β would be a k dimensional vector. If a particular β_i larger than another $beta_j$, then a Dirichlet distribution will have relatively higher likelihood assigned to distributions with higher probabilities assigned to category i than category j. The density of the Dirichlet distribution for probabilities $x_1, \ldots, X_k$ such that $\sum x_i = 1$ is given

DISTRIBUTION A: Distribution approved for public release.

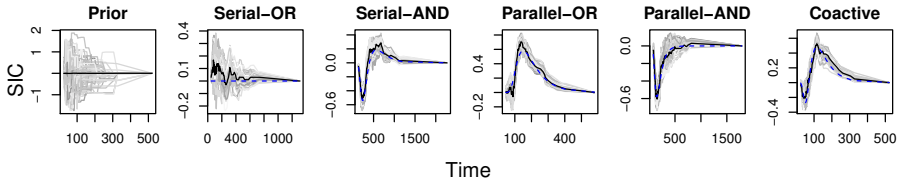

Figure 9: Sample SIC shapes from the DP model. The light gray lines represent individual samples from the prior/posterior distribution (100 of which are plotted) and the dark line represents the mean of the samples. For each of the five models, the SIC of the model that generated the data is plotted with a dashed line. These particular data are from the inverse Gaussian simulation study.

by,

$$f(x_1, \ldots, x_k; \beta) = \frac{1}{\mathrm{B}(\beta)} \prod_{i=1} k x_i^{\beta_i - 1}$$

where B() indicates the Beta function.

The Dirichlet process is an infinite dimensional generalization of the Dirichlet distribution, i.e., where k is infinite. In our application, it is a distribution such that for any binning of the RTs, the probability of a RT falling in each bin has a Dirichlet distribution. The Dirichlet process is parameterized by a function β which maps sets to positive real numbers,[5] β evaluated on the bins determines the β vector which parameterizes that Dirichlet distribution given by a particular binning of the RTs.

From a generative perspective, a random distribution is sampled from the Dirichlet process prior and the response times are randomly sampled from that random distribution. If we let $\mathrm{RT}_{l(i)}$ be the ith RT in condition $l \in \{\mathrm{HH}, \mathrm{LH}, \mathrm{HL}, \mathrm{LL}\}$, then,

$$\alpha_l \sim \mathcal{DP}(\beta)$$
$$\mathrm{RT}_{l(i)} \sim \alpha_l.$$

Samples from a Dirichlet process are almost surely discrete distributions whereas RTs are usually modeled with continuous distributions (e.g., Heathcote, 2004; Van Zandt, 2002). Nonetheless any continuous distribution can be approximated arbitrarily well by a discrete distribution, so we can obtain good approximations of continuous distribution with samples from the Dirichlet process prior.

For the application of the Dirichlet process prior to SICs, we must first consider the discretization of the RTs, i.e., how the RT bins will be defined. We did so using quantiles of the maximum likelihood inverse Gaussian fit to all of the RT data. This creates a compromise between two competing constraints, namely that the binning should be independent of the data and we would like to preserve as much information in the RTs as possible when binning the data.

[5] β must satisfy the formal definition of a measure, but is not required to be a probability measure.

DISTRIBUTION A: Distribution approved for public release.

In earlier work, Heathcote, Brown, Wagenmakers, and Eidels (2010) found little effect of violating the former constraint by choosing bins based on the quantiles estimated directly from the data.

Next, an appropriate β must be chosen. In our current approach, we fix β to be a uniform distribution across the defined bins.[6] This means that the relative probability that each bin is assigned a particular probability in a sample is the same across bins. Each of the four distributions used to calculate the SIC have the same prior, so for determined values of the prior, such as the mean or specific quantiles, the SIC is zero. However, because the samples from the prior for each distribution are random, samples of non-zero SICs from the prior are expected.

Computation of the Bayes factor simply involves counting the number of prior and posterior samples that meet the conditions that define a particular form of the SIC function. This approach is hampered by the fact that a flat SIC for the serial-OR model essentially has probability zero in the prior. Hence, if we had a continuous distribution over SICs, the flat SIC would have probability exactly zero. Because we are using discrete approximations, there is positive probability of a flat SIC, but it is small, and the better the binning approximates a continuous distribution, the lower that probability. This problem is not unique to nonparametric Bayesian statistics. Attempting to assess the probability of a point hypothesis (such as the null hypothesis that the means of two continuous random variables are the same) leads to the same issue.

To address this problem, we use a region of practical equivalence (ROPE) around zero (cf. Klugkist, Kato, & Hoijtink, 2005; Kruschke, 2011) for both the SIC and, for the Coactive/Serial-AND discrimination, the MIC (recall this is the integrated SIC). The ROPE is comprised of the values that are treated as zero, both in the prior and posterior distributions. Hence, any sample from a distribution over SICs that never falls outside the ROPE is treated as a flat SIC, any sample that extends beyond the positive bound of the ROPE, but not the negative bound, is treated as an all-positive SIC, and any sample that extends beyond the negative bound of the ROPE, but not the positive bound, is treated as an all-negative SIC. Similarly, if the MIC is within the ROPE, it is treated as zero. The range of the ROPE for the SIC need not be the same as the ROPE for the MIC, hence both are parameters that must be set for the model. We use values chosen based on a range of simulated models in the sections below. Ideally the test would not require the setting, or tuning, of specific parameters. Although this is considered a disadvantage in the realm of Bayesian statistics, it is similar in practice to choosing an alpha level for a null-hypothesis-significance test.

To determine the relative evidence among the potential SIC forms we used BFs calculated using the encompassing prior approach (Klugkist et al., 2005). First we estimate the posterior probability of each SIC shape. We do so by sampling distributions for each of SIC conditions from the prior and from the

[6] To build a hierarchical model over the different RT distributions, one could also assign a separate β to each salience condition, each of which is drawn from another Dirichlet process.

DISTRIBUTION A: Distribution approved for public release.

posterior, then counting the frequency with which those samples yield an SIC that is 0 (within the ROPE for all bins), all negative, all positive, negative-then-positive with MIC within its ROPE, negative-then-positive with positive MIC, or something else (e.g., positive then negative). The BF for a particular SIC shape is then estimated by the ratio of the number of samples in the posterior and prior. Details on implementing this approach are included in the supplementary material.

To obtain estimates of the prior and posterior probability of any particular SIC form we examine the combination of the priors and posteriors for each condition's RT distribution. Examples of samples from the prior and posterior of the DP model of the SIC shape are shown in Figure 9. The majority of the prior probability does not correspond any of the SIC forms of interest (e.g., there are many positive and negative ranges), approximately 0.729.

Although the DP test allows for the possibility of SIC shapes other than those depicted in Figure 1, we give those other shapes zero prior probability. This is done in practice by ignoring samples from either the prior or posterior that are not classified as one of the five predicted shapes. Hence, BFs compare a given model to an encompassing model constituted of the union of the five SIC shapes.

When conditioned on having one of the five forms corresponding to the generating models of interest, all-negative and all-positive have by far the highest prior probability (both approximately 0.49) while the coactive form has the next highest (0.0017) followed by Serial-AND (1.95×10^{-5}) and Serial-OR (5.16×10^{-7}). While the encompassing BF approach can account for the differential prior weighting, the extremely low probability of a Serial-OR form can lead to poor prior estimates unless very large samples are taken for the prior and posterior. To compensate, we added a single sample to any category for which there was no observed sample, both for the prior and posterior. When the number of samples for each is set to $10,000$, as it was in all of our analyses, this effectively sets the minimum probability of a SIC shape to 1×10^{-4} in both the prior and posterior.

3.2 Hierarchical Bayesian Ideal Observer Analysis[7]

In many cases, the more information or higher quality information an individual has, the better decision that person can make. The exact mapping between information quality and decision is frequently the main focus in cognitive and perceptual research. A deeper, and potentially more revealing question about the nature of human perception and cognition is the efficiency with which an individual can use available information. Furthermore, it is often informative to compare how efficiently individuals use different types of information. To do so, it is necessary to account for the amount of information across two potentially disparate types of information. One of the more successful methods for making these comparisons is ideal observer analysis, which has been used

[7]The content from this section is from Houpt and Bittner (n.d.).

DISTRIBUTION A: Distribution approved for public release.

most prolifically in studies of human visual perception and psychophysics. Well known applications include 3D vision (e.g., Liu & Kersten, 2003; Tjan, Braje, Legge, & Kersten, 1995), symmetry (e.g., Barlow & Reeves, 1979; Liu & Kersten, 2003), face perception (e.g., Gaspar, Bennett, & Sekuler, 2008; Gold, Bennett, & Sekuler, 1999; Gold, Sekuler, & Bennett, 2004), biological motion (e.g., Gold, Tadin, Cook, & Blake, 2008), and many others (for a review, see Geisler, 2011).

Ideal observer analysis gives a direct measure of a human observer's ability to use available information without the confound of information content by formalizing the construct of *efficiency*. This is done through a comparison of human performance to that of an optimal decider (i.e. ideal observer) that uses all information available in an experiment. The calculation of human efficiency based on ideal observer analysis is valuable to psychological research as it provides a measure of human information processing that can be examined across experimental conditions while accounting for the variation in information due to the condition itself. This allows for human efficiencies to be directly compared across wide ranges of tasks, stimulus types, and other experimental manipulations.

With the flexibility and the increased popularity of ideal observer analysis, a progression of research examining tasks and stimuli of increasing complexity has arisen over time (see Geisler, 2011). Naturally, this has led to efficiency results containing attributes of data not seen or considered in the original application of the techniques including learning, (Gold et al., 1999, 2004, e.g.,), and individual differences (Brainard, Williams, & Hofer, 2008, e.g.,). The current paper addresses the statistical issues that can accompany such increased complexity of ideal observer applications. Specifically, we provide solutions for addressing variability in data that can arise from differences among individuals and experimental manipulations, as well as uncertainty in estimates of ideal performance. We demonstrate a hierarchical Bayesian model that can be used for inference based on individual and group level efficiency. We then demonstrate our approach applied to both a simulated dataset and data from human experimentation examining multispectral image fusion.

To calculate efficiency, we must first understand how much information is available in a given experimental test. In ideal observer analysis, a computationally optimal decider (i.e. ideal observer) that uses all of the relevant information available to complete a task is derived. Hence, variation in ideal observer performance across experimental manipulations indicates variation in the amount of relevant information (e.g., changes to stimuli and/or task specifications). This allows ideal performance to serve as a benchmark for comparison with human data to account for potential variations in information. To compare across stimuli, tasks, or other information content, the ratio of human to ideal performance can be used. This ratio gives a measure of what percentage of the available information the human utilizes (i.e. human efficiency). Comparing efficiencies across experimental conditions gives a direct measurement of information use, allowing for a deeper understanding of the human perceptual system.

There are number of important considerations relating to task design and performance measurement when applying ideal observer analysis: (1) The spe-

DISTRIBUTION A: Distribution approved for public release.

cific aspects of the experimental stimuli and task paradigm, (2) the metric by which performance is evaluated, (3) derivation of the ideal observer decision rule and computation, (4) calculation of efficiency, and (5) comparison of efficiency values across the items of study. Each of these aspects require definition driven by the theoretical aspects of the topic at hand and/or the inherent definitions of optimality and information.

First, the experimental design and parameters must be well specified. This constraint is important for all psychological studies and it is particularly important for ideal observer analysis because the specifics of the task and stimulus information available are critical to the derivation of ideal performance. Specifically, the ideal observer must have "knowledge" of all experimental components. For example, within perceptual experiments, this knowledge is usually encoded as a separate template representing the expected percept for each possible stimulus. (This follows from the fact that template matching leads to optimal performance in many perceptual experiments.) Additionally, the ideal observer must have a representation of the task (i.e. what and how information is being compared throughout the experiment).

The second consideration is the metric by which performance of the human and ideal observer are assessed. There are two main strategies used for this approach. One is to use a traditional accuracy or accuracy related measure, such as percent correct, hit rate, or d' (a measure from signal detection theory meant to remove observer bias from the performance metric, e.g., Macmillan & Creelman, 2005; Tanner Jr & Birdsall, 1958). Alternatively, the performance level can be fixed and the stimulus can be varied to determined the level of stimulus variation for the observer to reach the level of performance, a metric referred to as the *threshold*. The latter metric is the most commonly used in recent applications of ideal observer analysis and we demonstrate our analysis model in terms of thresholds, although our suggested analysis applies equally well to the former class of performance criteria. Whether accuracy or thresholds are used, efficiency is defined as the ratio of the human performance to the ideal performance, e.g., efficiency for thresholds is the ratio of the ideal observer's threshold to the human observer's threshold. For accuracy or d' ratios, the ideal observer's performance is used in the denominator because higher d' or accuracy implies better performance. For threshold measures, higher values imply worse performance (more is needed to achieve the criterion level of accuracy) so the ideal observer's threshold is used in the numerator. Either way, the maximum possible efficiency is 1 (or 100%), i.e., performance equal to the ideal observer.

Efficiency can not be estimated from a single trial. To establish a human performers accuracy, a number of trials will need to be repeatedly presented. To estimate a threshold, samples of human observer performance are needed with a variety of stimulus levels. There are a number of approaches to choosing what to present to an observer on any given trial with varying methodological sophistication. Perhaps the most straightforward is the method of constant stimuli, in which a predetermined set of stimulus levels is used regardless of observer performance (e.g., Urban, 1910). Alternatively, adaptive methods can be used, in which stimulus level is varied as a function of observer performance. This

DISTRIBUTION A: Distribution approved for public release.

class of designs including staircase procedures (e.g., Békésy & Wever, 1960), in which variation is according to recent errors and correct responses, and QUEST (A. B. Watson & Pelli, 1983) and Ψ (Kontsevich & Tyler, 1999) which use Bayesian updating and a maximum informativeness criterion to determine the level of stimulus to present. Regardless of the method used to choose stimulus levels, the most effective way to determine the threshold is by fitting a model to the pattern of results. This model fitting is done by varying the parameters of a function that maps stimulus level to accuracy, i.e, the psychometric function (cf. Figure 10). We focus on this model based approach to estimating thresholds because not only does it determine a threshold value with consideration for data collected on every trial but it naturally allows for calculating the likelihood of the data (see Gold et al., 2008).

Third, the definition of an ideal decision requires a formal statement of the optimal decision strategy, i.e., the strategy that accounts for all experimental procedures and constraints. In the traditional analysis, the ideal strategy for making probabilistic comparisons defined through stimulus and task constraints to produce responses across experimental trials is derived using Bayes' rule. In perceptual studies, the optimal strategy is based on the maximum a posteriori template given a stimulus. It is important to note that, without some uncertainty (i.e., with imperfect information) the ideal observer will always give the correct response because the stimulus would have probability 1 while all alternatives would have probability 0. In the majority of applications of ideal observer analysis, stimuli are presented with zero-mean Gaussian white noise added. Examples of decision rule derivations for various tasks and noise types can be found across ideal observer literature (e.g., Liu & Kersten, 2003; Tjan et al., 1995) and will not be covered here in detail. The interested reader is encouraged to explore these specifications in such literature.

After obtaining efficiencies across experiments or experimental conditions, a statistical comparison of the data is necessary to derive conclusions about the constructs being studied. With any performance value estimation, there is necessarily some amount of inherent uncertainty. Any statistical inference requires assumptions about that uncertainty. For example, the traditional application of ANOVA to efficiency assumes that the variation in performance across observer and/or condition is large relative to the uncertainty in the threshold estimates of the human and ideal observer. In many of the previous applications of ideal observer analysis, researchers have studied a small number of observers and used a large number of trials for precise individual threshold estimates. For applications such as these where a large number of trials are used to estimate each threshold, this assumption is reasonable. Indeed, the ideal observer threshold can often be estimated with arbitrary precision because the only cost with achieving better estimates is computational time. However, human observers' time is often more costly, particularly for special populations such as experts, so arbitrary precision is unattainable.

When there is a consequential amount of uncertainty in either the human or ideal observer performance estimates, it must be appropriately addressed to make valid inferences. Furthermore, regardless of the empirical and estimation

DISTRIBUTION A: Distribution approved for public release.

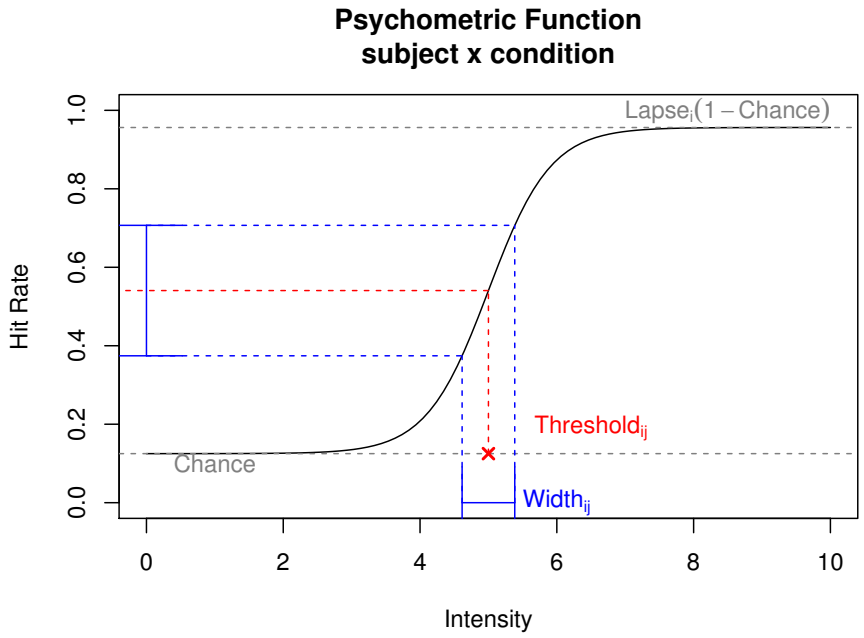

Figure 10: Example psychometric function that maps signal strength (e.g., intensity, contrast, etc.) to hit rate. The lower bound of performance is given by chance guessing (in this case 0.125 and the upper bound of performance is limited by the frequency of attentional lapses. Psychometric functions are usually specified with two additional parameters, the threshold (i.e., the halfway point between minimum performance and maximum performance) and the width (change in intensity from the threshold that results in a prespecified changes in hit rate). In general, chance performance is determined by the task, while lapse rates can vary across observers and thresholds and widths can vary with condition and observer.

DISTRIBUTION A: Distribution approved for public release.

choice, thresholds are traditionally estimated independently for each observer in each particular condition. Group or population level inferences as well as the global effects of the variation in condition which are often assessed as a follow-up to threshold estimation.

Another issue of concern is distribution of the uncertainty. Standard ANOVA techniques assume normally distributed uncertainty which is not necessarily a reasonable assumption for data derived from accuracy estimates or ratios of estimated variables. One solution to this is to use nonparametric variants of ANOVA, such as the Friedman test (Friedman, 1937).

Alternatively, we can leverage our beliefs about the generating processes of interest to estimate the distribution of uncertainty. By using Bayesian analyses we are able to to incorporate prior knowledge and use what we know about the structure of the data to infer the appropriate distributions.

A further advantage of Bayesian data analysis is the relative ease with which to implement a hierarchical analysis. This allows for both individual and group level inferences regarding efficiency and can improve inferences in the presence of outliers due to shrinkage.

In this section we present an approach to estimating posterior distributions of psychometric functions and particularly efficiency levels across individual human observers and experimental conditions. The model we outline is specific to efficiency measures estimated from ratios of thresholds, but a model based on ratio of d's or accuracies would follow a similar form (see also Rouder & Lu, 2005).

The first step in developing the model of the psychometric functions and thresholds is to determine the appropriate likelihood function for the observed data. With accuracy data (or any series of dichotomous outcome data) a common Bayesian modeling approach would be to model the outcome as an observation of a binomial random variable. One standard frequentist approach to analyzing accuracy data is to assume (sometimes implicitly) that the mean accuracy for each treatment is normally distributed (sometimes after a transformation of the raw accuracies) and then use an ANOVA to test for significant effects of the treatment. A more sophisticated approach is to assume that the outcome variable from the ANOVA model is transformed to the appropriate scale via a logistic transform. Indeed, this approach is at the heart of our currently proposed analysis, although we use a Bayesian version (cf. Kruschke, 2010, Chapter 21) for the reasons summarized in the introduction.

Thus, we will assume that each response is a sample of a Bernoulli random variable and hence a sequence of trials within a given treatment is distributed as a Binomial random variable. For a given trial, the only free parameter of a Bernoulli random variable is the probability of a success (e.g., a hit) which is given by the psychometric function evaluated for the signal intensity on that trial. Our basic model of the psychometric function essentially follows the modeling framework from Kuss, Jäkel, and Wichmann (2005).

Recall that the psychometric function, $\Psi(\cdot)$, maps the stimulus intensity to the probability of a correct response. The standard approach in psychophysics is to assume that the $\Psi(\cdot)$ is some monotonically increasing function determined by

DISTRIBUTION A: Distribution approved for public release.

a small number of parameters. In most applications, the limit of the psychometric function as the stimulus intensity goes to zero is the accuracy rate predicted by uniform random guessing (π_c, e.g., if there are four possible responses, then the accuracy limit as the stimulus intensity goes to zero is $\pi_c = 0.25$). As such, this is not usually treated as a free parameter, but is instead fixed by the properties of the experiment. The upper limit of Ψ, accuracy as stimulus intensity is maximized, π_l, is governed by a free parameter. This limit is the frequency with which an observer makes an incorrect response despite have the clearest possible information, which are generally explained by attentional lapses which are not determined a priori. Although we do not know the exact value of π_l, we can build some vague knowledge about it into the model. For example, experimenters will often not bother to analyze data from human observers that are not paying attention for large portions of the experiment, so we may want to assume that the lapse rate is small, e.g.,

$$\pi_l \sim \beta(1, 99).$$

The next parameter of interest, perhaps of most interest, is the location parameter, m, which gives the intensity level at which the accuracy is 50% in the absence of guessing or lapses. If we are only interested in a single observer in a single condition, then it is sufficient to assume some prior distribution on m based on the intensity scale of interest. Unlike the lapse rate, the a priori values for m can be wildly different across different experiments (e.g., possible thresholds for volume can be totally different than possible thresholds for brightness). Note that, in addition to the location and scale of the prior distribution, the shape may also be different across scales. For example if a symmetric (e.g., normal) distribution is reasonable for intensities on a log-scale (cf. Weber's law) then an asymmetric (e.g., log-normal) distribution is correspondingly reasonable on the raw scale. Furthermore, 0 is usually the lower limit on raw intensity scales, and when that is the case, the prior should reflect that constraint. For example, we can a priori rule out the possibility that a persons threshold volume for detecting a tone is negative. As we will see below, there are some cases in which one may want to include more in the model for m, particularly how m varies across treatments, learning (e.g., Gold et al., 2008) and individual differences (e.g., Brainard et al., 2008).

Next is the width, w, parameter (or similarly the slope or scale parameter) for Ψ. This parameter controls the rate at which the accuracy increases as a function of increasing stimulus intensity. Following Kuss et al. (2005), we focus on w as the difference between the intensity level that would result in 10% accuracy and that which would result in 90% accuracy if chance performance and the lapse rate were 0. Although this parameterization is not crucial to the further development of our model, the fact that it is in terms of the intensity scale makes it more straightforward to assign meaningful priors.

There are essentially infinite degrees of freedom in determining the exact function with location m and width w that maps from intensity to accuracy in the range between π_l and π_c. Kuss et al. (2005) cover Bayesian implementations

DISTRIBUTION A: Distribution approved for public release.

of the more common functions that are used in this role including the logistic, probit, Gumble and reverse Gumbell and Weibull and Reverse Weibull. Usually one of these functions are sufficient, but other functions may be useful depending on the properties of the psychophysical scale. For the remainder of this paper, we use the logistic function because of its connection to logistic regression. Using the m, w parameterization with x indicating intensity,

$$F(x) = \left[1 + \exp\left(\frac{-4.39(x-m)}{w}\right)\right]^{-1}. \tag{9}$$

The full psychometric function mapping intensity (x) to accuracy is given by,

$$\Psi(x) = (1-\pi_l)\left[(1-\pi_c)F(x) + \pi_c\right] + \pi_c\pi_l. \tag{10}$$

Verbally, this equation can be interpreted as follows: If the observer does not lapse, $(1-\pi_l)$, then she either does not guess, $(1-\pi_c)$, and hence uses the signal to the best of her ability, $F(x)$, or she guesses, (π_c). If she does lapse, then she guesses, $(\pi_c\pi_l)$.

Given a mapping between the threshold, m, and width, w, of the psychometric function, we now need a way to represent the changes in those parameters across conditions and observers. Here we follow the assumptions of an ANOVA, that each parameter is given by the grand mean μ, some deviation from that mean specific to a condition, α, some deviation specific to the observer γ and, potentially, deviation due to observer-condition specific interactions (see Figure 11),

$$m_{ij} = \mu^{(m)} + \alpha_j^{(m)} + \gamma_i^{(m)} + (\alpha\gamma)_{ij}^{(m)} \tag{11}$$

$$w_{ij} = \mu^{(w)} + \alpha_j^{(w)} + \gamma_i^{(w)} + (\alpha\gamma)_{ij}^{(w)}. \tag{12}$$

In practice, to be certain the μ represents the grand mean across all samples in an MCMC chain, we force $\sum_j \alpha_j = 0$ and similarly for the other parameters. It is important to note that, while this seems similar to the assumption of fixed effects in an standard ANOVA, it does not have the same interpretation within the Bayesian framework. Within the Bayesian approach, all parameters with distributions are random variables and hence effectively representing random effects. The difference between drawing conclusions restricted to the observed αs and drawing conclusions about any possible α is instead based in how the posterior is used.

Setting priors for each of the parameters in Equations 11 and 12 again depends on domain specific knowledge, however some aspects of the prior can be determined regardless of the domain. If the model is used with a raw intensity scale, then m must be positive, so $\mu^{(m)}$ must be positive and,

$$\min_{ij}\left(\alpha_j^{(m)} + \gamma_i^{(m)} + (\alpha\gamma)_{ij}^{(m)}\right) < -\mu^{(m)}.$$

If a log-intensity scale is used, then it is sufficient to require m to be finite, which in turn only restricts the other m parameters to being finite. The width

DISTRIBUTION A: Distribution approved for public release.

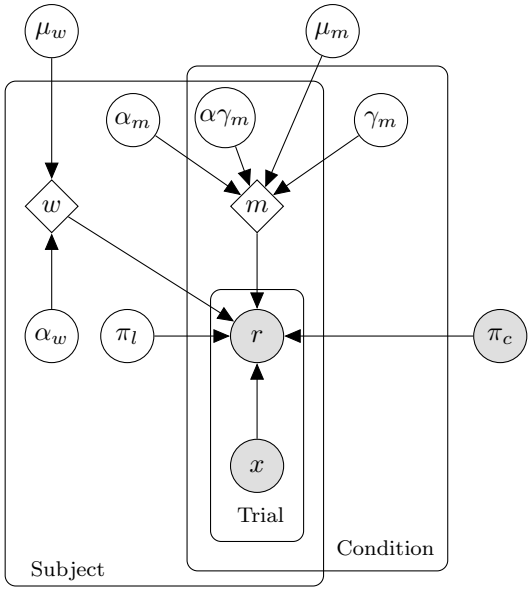

Figure 11: Diagram of the hierarchical structure of the Bayesian model. Shaded nodes indicate observed variables. Unfilled circular nodes indicate parameters that are random variables. Diamond nodes indicate parameters that are determined by their parents. The innermost plate indicates there is a signal intensity, x, and response r, for each trial. The probability of the response is affected by chance performance, π_c, which is a property of the experiment and hence fixed across trials, subjects and conditions. Each trial specific to a particular subject and condition. Across conditions, each subject has their own lapse rate, π_l, and psychometric function width w which is determined from a group level prior mean μ_w and individual deviation from that mean α_w. For each subject in each condition, there is a threshold parameter, m, determined by a group prior μ_m, deviation due to subject α_m, condition, γ_m, and an interaction between subject and condition $\alpha\gamma_m$.

DISTRIBUTION A: Distribution approved for public release.

must be positive, regardless of whether the intensity is raw-scale or log-scale, so $\mu^{(w)}$ must be positive and

$$\min_{ij}\left(\alpha_j^{(m)}+\gamma_i^{(m)}+(\alpha\gamma)_{ij}^{(m)}\right)<-\mu^{(m)}.$$

Prior distributions on α, γ, and $(\alpha\gamma)$ should all be zero centered as they represent deviation from the grand mean. Assuming some form of Weber's law holds for the intensity, then the distributions for the m parameters should be positively skewed on the raw-scale or symmetric on the log-scale to reflect that at smaller intensities, the perceptual scale is compressed. There is less prior reason for the w parameters to be positively skewed (other then the aforementioned requirement that $P(w>0)=1$).

The variability of the prior distribution for α reflects the magnitude of difference a change in condition might make on the threshold. Similarly, the variability of the prior on γ reflects the prior belief in how different individuals' thresholds will be averaged across conditions and the variability of the prior on $(\alpha\gamma)$ represents the a priori belief in how much change from the main effects is expected in a given combination of observer and condition. In all cases, the default is to assume a fairly diffuse prior (i.e., high variance) to correspond with the assumed prior knowledge in a traditional ANOVA.

As the reader may have intimated, there is no reason to constrain the model to a single factor such as the condition. For example, if one were interested in changes across conditions and across multiple days of testing, the same mapping from the ANOVA model to the current framework applies, i.e,.

$$\begin{aligned}
m_{ijk}&=\mu^{(m)}+\alpha_j^{(m)}+\beta_k^{(m)}+\gamma_i^{(m)}+(\alpha\beta)_{jk}^{(m)}+(\alpha\gamma)_{ij}^{(m)}+(\beta\gamma)_{ik}^{(m)}+(\alpha\beta\gamma)_{ijk}^{(m)}\\
w_{ij}&=\mu^{(w)}+\alpha_j^{(w)}+\beta_k^{(w)}+\gamma_i^{(w)}+(\alpha\beta)_{jk}^{(w)}+(\alpha\gamma)_{ij}^{(w)}+(\beta\gamma)_{ik}^{(w)}+(\alpha\beta\gamma)_{ijk}^{(w)}.
\end{aligned}$$

Likewise, the same linear model can be used with continuous independent variables in a linear regression type approach, i.e., with X representing the level of the independent variable,

$$\begin{aligned}
m_{ij}&=\mu^{(m)}+\beta^{(m)}X_j+\gamma_i^{(m)}+(\beta\gamma)_i^{(m)}X_j\\
w_{ij}&=\mu^{(w)}+\beta^{(w)}X_j+\gamma_i^{(w)}+(\beta\gamma)_i^{(w)}X_j.
\end{aligned}$$

With a complete model of the psychometric function, posterior distributions of any level threshold can be obtained, whether it be the 50% threshold (i.e., m) or any other level. In the next section, we describe how these posterior distributions can then be used to obtain estimates of the posterior distribution of efficiency.

The main impetus behind ideal observer analysis is to derive human efficiency. This construct is determined through the ratio of human to ideal performance. The purpose of this derivation is to provide a measure of human performance that reflects information use in the visual system without the confound of experimental information. Raw performance measures reflect human

DISTRIBUTION A: Distribution approved for public release.

ability within an experiment, however they do not account for inherent information in the study itself. Efficiency provides a direct measurement of information usage the can be compared and contrasted across tasks, stimuli, and other such experimental manipulations.

Hence, for Bayesian inference based on efficiency, both a posterior over the ideal observer's threshold and the human observers' thresholds are needed. Assuming ideal observer's responses are conditionally independent of the human observer's responses given the condition, we can obtain posterior samples for efficiency by dividing posterior samples of human (individual or group level) thresholds by posterior samples of ideal observer thresholds. In particular, if we are interested in posterior estimates of the group level efficiency in Condition 1 , then we would combine MCMC samples $n = \{1, \ldots, N\}$ of the group mean location for the human observers ($m_{\cdot j}^{(n)}$) with MCMC samples of the ideal observer in that condition $m_{\mathrm{IO}j}^{(n)}$,

$$e_{\cdot j}^{(n)} = \frac{m_{\cdot j}^{(n)}}{m_{\mathrm{IO}j}^{(n)}}. \tag{13}$$

Samples from the Equation 13 can then be combined to obtain an Monte-Carlo estimate of the posterior distribution of $e_{\cdot j}$. The same principal holds for other potential efficiencies, such as for individual observers or specific condition–observer combinations.

To better understand the behavior of the analysis, we ran a simulation study that included a group level effect of condition as well as individual variation. Data were simulated using the model described in Equations 10, 11 and 12. The grand mean threshold, $\mu^{(m)}$, was set to 5 and the grand mean width, $\mu^{(w)}$, was set to 3. The effect of the condition on thresholds, $\alpha_j^{(m)}$ were randomly generated from a normal distribution with mean 0 and standard deviation 1. We did not include any effect of condition on the width parameter. Individual subject effects were generated from a normal distribution with mean 0 and standard deviation 0.1 for the threshold and standard deviation 0.05 for the width. Finally, each observer's lapse rate was sampled from a Beta distribution with $\alpha = 5, \beta = 95$. For observed data, we generated correct and incorrect responses from Ψ following a method of constant stimuli, with 60 samples each at intensities $3.5, 4.42, 5.28, 6.13$ and 6.88 for each observer. We simulated a within-subject design with 4 conditions and 10 subjects.

Prior and posterior distributions for the condition parameters and the subject parameters are shown in Figures 12, 13 and 14. The "X" indicates the value used to generate the data. Posterior distributions for the condition parameters are quite tight and tend to be close to the parameter used to generate the data. The subject-threshold posteriors were more variable and not necessarily centered on the generating parameters. This result makes sense given that there were more observations per condition than observations per subject. Subject width posteriors did not vary from the prior, indicating there was not strong evidence in the data for determining the exact value of the width. This

DISTRIBUTION A: Distribution approved for public release.

finding agrees with previous comparisons between adaptive experimental design approaches that indicate more trials are needed to achieve precise estimates of the width parameter than the threshold parameter (e.g., Kontsevich & Tyler, 1999).

To demonstrate the strength of our technique, we apply our process to two experimental datasets. Both sets come from studies that examine image fusion, a multispectral image enhancement involving the algorithmic combination of two single-band image components. In each experiment, human efficiencies were collected over 9 blocks of trials: 2 blocks corresponding to trials using single-band spectral imagery and 7 blocks corresponding to 7 different image fusion combinations. We varied contrast to manipulate the amount of information available in a stimulus and added Gaussian noise to an image before displaying it. Human and ideal performance was measured by determining the threshold contrast in the image necessary to achieve 50% accuracy (71% in the shove condition because chance performance was higher). The data were collected using a staircase procedure. The goals of each study relevant to our demonstration were to determine if (1) human efficiency is better in fused conditions blocks over single-band blocks and (2) are there efficiency differences between image fusion blocks? Similar to the simulated data set, we are interested in both group-level inferences and an understanding of individual variation.

The two datasets differ in their image components and tasks. The first set utilized a 1-of-8 orientation task using Landolt C images (Figure 15) and the second set contained data from a 1-of-2 task of two scene images with an actor holding a shovel on the left or right hand side of his body (Figure 16). Landolt C imagery was created using a white insulator with a cutout in front of a dark plate that could be heated to increase the LWIR signal. Both sets involved fusion of single-band imagery taken in the visible and long-wave infrared (LWIR, i.e. thermal) spectra. Data from the sets are shown in Figure 17.

One major difference in the example data lies in the variability in individual efficiencies. In both sets, observers vary in magnitude of efficiency values. This finding is typical in most ideal observer experiments. More importantly however is that the data for the set using the shovel task vary greatly in order of effects. Namely, individual human observer efficiency rankings between image sets are not consistent. This is not the case in the Landolt C data. Here, observers data generally follow similar patterns in efficiency ranking of image sets.

Another notable feature of our approach is the ability to include all human observer data. Under many traditional threshold estimations, observers who perform at low accuracy levels with more information produce patterns of decisions that do not allow for reliable threshold estimation. Thus, the data has to be discarded in the final analysis. Here, this data is preserved in the full analysis by using the hierarchical structure to constrain possible psychometric function parameters and hence be able to fit the lower accuracy observers.

The efficiency data for the shovel task show primarily comparable results between image sets, with a potential for single-band thermal data imagery to be producing a stronger impact on efficiency. In the Landolt C data, there also appears to be a potential for single-band imagery to be positively affecting ef-

DISTRIBUTION A: Distribution approved for public release.

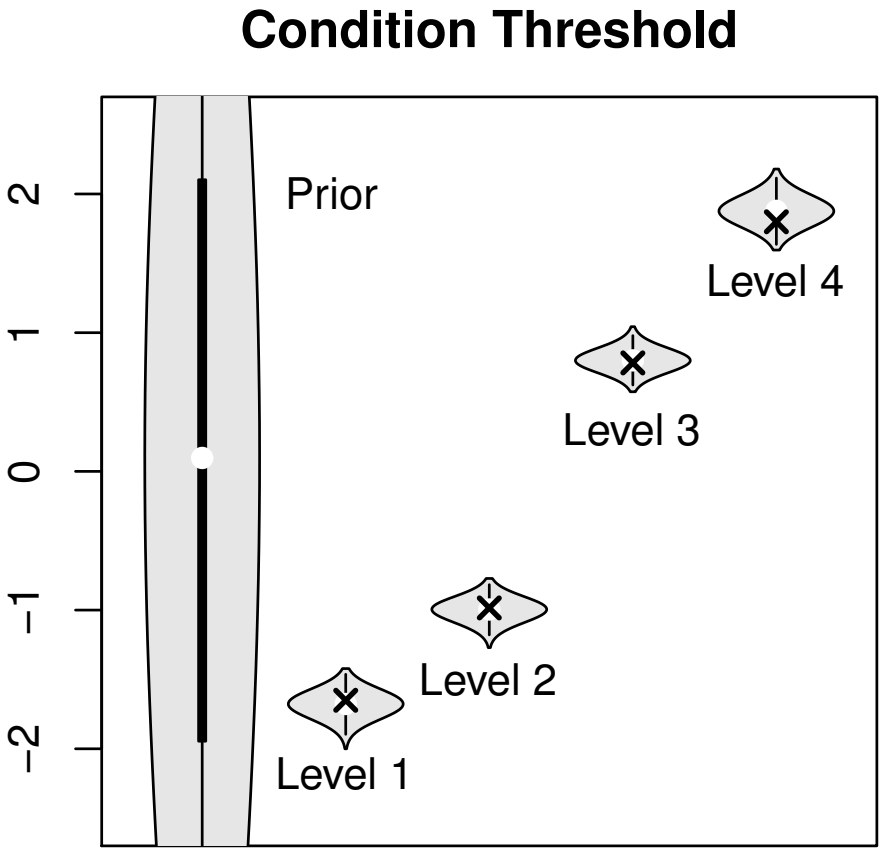

Figure 12: Prior and posteriors for condition effect on threshold of psychometric function from simulated data. The "X" indicates the parameter used to generate the data.

DISTRIBUTION A: Distribution approved for public release.

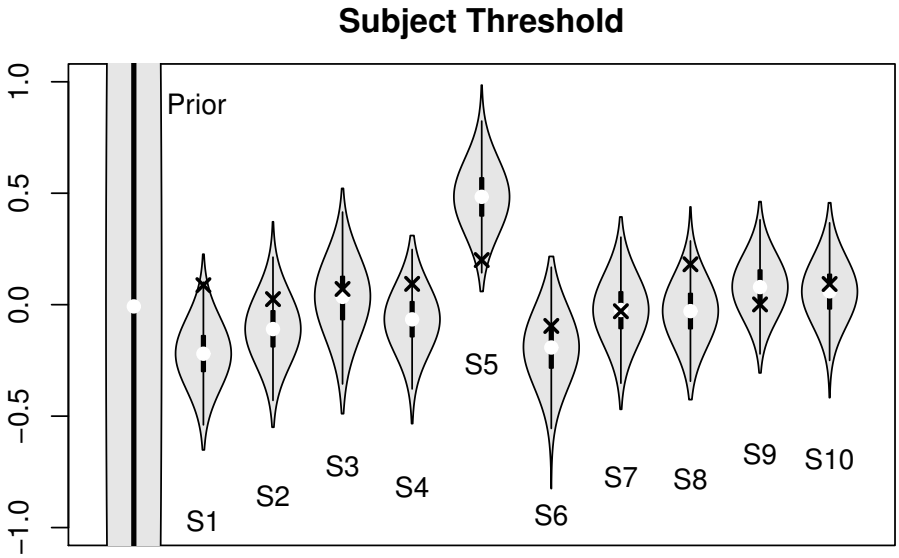

Figure 13: Prior and posteriors for subject effect on threshold of psychometric function from simulated data. The "X" indicates the parameter used to generate the data.

DISTRIBUTION A: Distribution approved for public release.

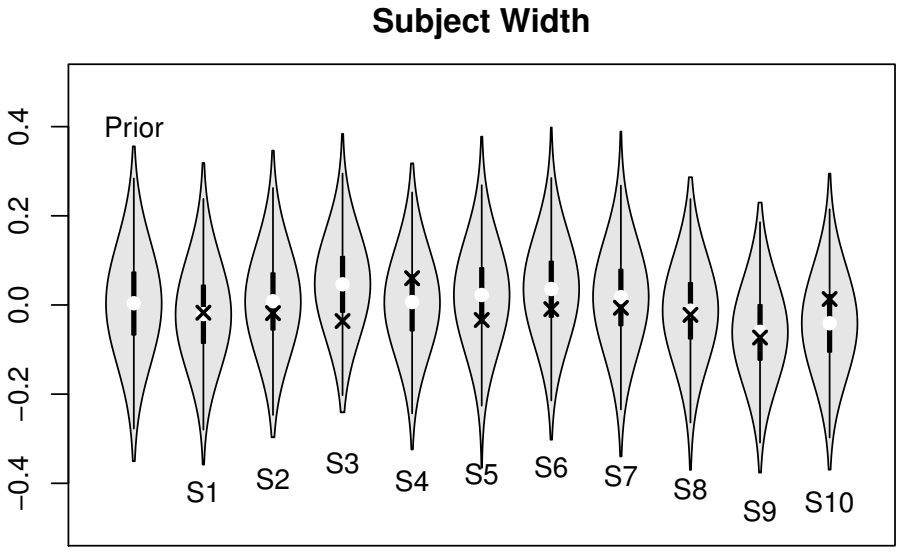

Figure 14: Prior and posteriors for subject effect on width of psychometric function from simulated data. The "X" indicates the parameter used to generate the data.

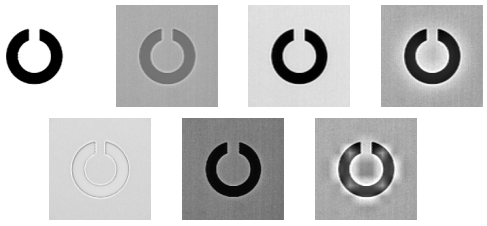

Figure 15: A visible and thermal Landolt C fused using various image fusion algorithms. From left to right they are: PCA, averaged, adjusted PCA, Lapacian pyramid, maximum, minimum, and wavelet pyramid. The details of these algorithms are beyond the scope of this paper, but we give the names for comparison to the posterior efficiency estimates.

DISTRIBUTION A: Distribution approved for public release.

Figure 16: Example Shovel-Task stimuli with both raw images and various fusion algorithms.

ficiency, this time in the visible case. However, there is an even more apparent pattern where the image fusion algorithm maximum produced extremely low efficiency values. The variability in individual data efficiency orderings, mentioned earlier, is also captured. When comparing between the two graphs, it is immediately apparent that the distributions on the shovel data are much more spread than those in the Landolt C data.

We have demonstrated a new method for analyzing human perceptual thresholds along with ideal observer thresholds using a Bayesian approach. We followed Kuss et al. (2005) in modeling each individual psychometric function with a likelihood based on the mixture of chance guessing and responding affected by the stimulus intensity. We used the logistic function, parameterized by its location (i.e., inflection point) and width (i.e., scale) although the theoretical approach applies equally to other standard functions that have been used to map intensity to accuracy. To model variation in performance across task condition and across individuals, we have leveraged the generalized linear modeling framework, taking inspiration from hierarchical Bayesian signal detection models from Rouder and Lu (2005) and Bayesian ANOVAs from Rouder et al. (2012). This led us to propose a linear combination of condition factors and subject factors as the core of the model, which are then used to determine the location and width of the logistic function. The linear core is quite flexible and can be used to model multiple factors and their interactions as well as either discrete factor levels (cf. ANOVA) and continuous factors (cf. regression).

The simulation demonstrated that with moderate numbers of samples, precise qualitative conclusions and reasonably good quantitative conclusions about

DISTRIBUTION A: Distribution approved for public release.

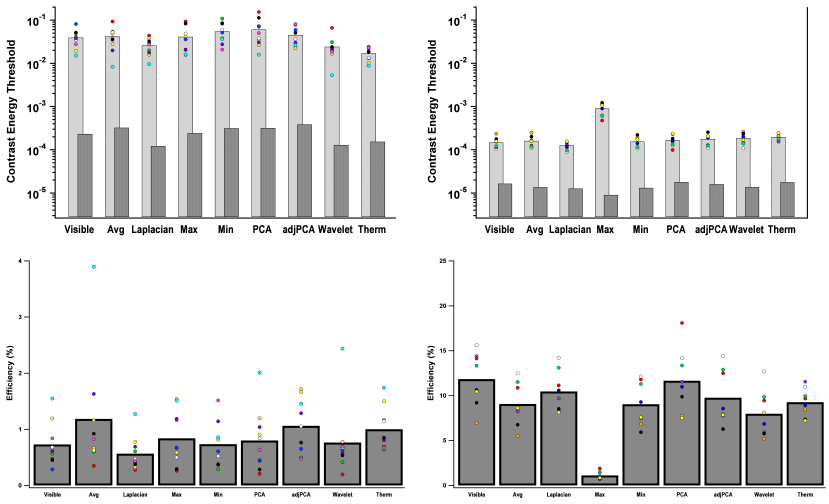

Figure 17: Top row: Maximum likelihood estimates of thresholds for individual human observers (points) and ideal observer (dark grey bar) along with group average threshold (light grey bar). Bottom row: Individual efficiency estimates based on the maximum likelihood thresholds (points) and group average efficiency (bar). The graphs on the left are the data from the Shovel experiment; on the right are the data from the Landolt C experiment. Note that not all observers from the Shovel experiment that were used in the Bayesian anlaysis are represented here because the maximum likelihood estimate failed to converge for some observers in some conditions.

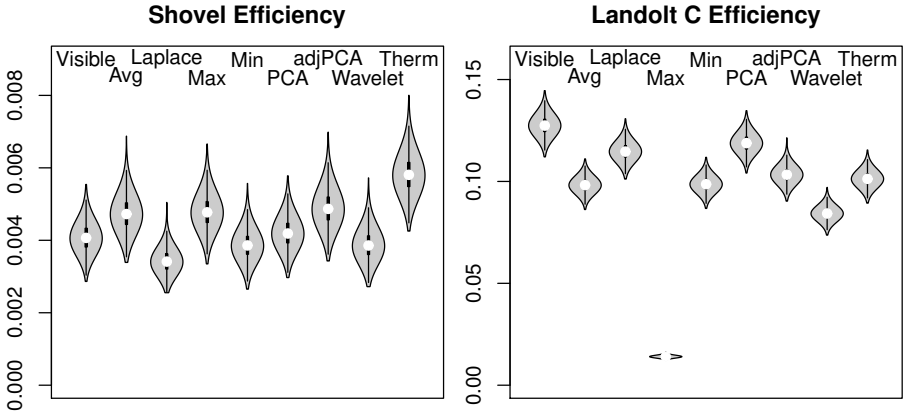

Figure 18: Posterior distribution of group level efficiencies for shovel imagery and Landolt C imagery. Individual sensor imagery (visible and thermal) are on the ends with the fused images using various algorithms in between.

DISTRIBUTION A: Distribution approved for public release.

the variation of thresholds (and hence efficiencies) can be drawn. Furthermore, in our application to recently collected data, we were able to obtain estimates of group level performance and a measure of the uncertainty of those estimates. These values can be used to inform decision making about the relative value of deploying these various fusion algorithms. Although we did not cover them here, estimates of individual performance and lapse rates were also obtained.

The use of our new technique provides valuable insight into threshold and efficiency estimations. With this tool, we are now able to obtain a deeper level of understanding of the impact of experimental manipulations upon human information usage. Specifically, the flexibility of our technique provides a more thorough look at the impact of experimental factors upon efficiency both within and between subjects. This process opens the gates for more complex datasets to be examined in a highly principled manner. With this approach, we can now meet the growing interest and expansion of ideal observer analysis studies.

4 Basic Research on Visual Search

4.1 Color and Shape in Visual Search

The first of our two experiments was designed to investigate the architecture, stopping rule, stochastic independence, and workload capacity of the cognitive system responsible for integrating color information and shape information in the context of single-feature visual search. While the search target was defined by the conjunction of two features, all the distractors on a given trial were identical, and thus on some trials the target differed from distractors by a minimum of one feature and maximum of two features.

The study was conducted at Wright State University and approved by its Institutional Review Board. Experiment 1 took place over four one-hour sessions; we did our best to ensure all four sessions took place within the period of one week and no more than two weeks.

Participants. All participants gave written informed consent before beginning the study and were compensated $8 per one-hour session. Participants who completed all four sessions were awarded an additional $2 per session bonus (i.e. $40 maximum compensation). Eighteen participants began the experiment, seventeen of whom completed every session. Additionally, data from two participants were discarded for exceptionally poor accuracy, suggesting that they did not understand the task. Data from fifteen participants remained for analysis.

Materials. Experiments were developed using PsychoPy (Peirce, 2009) libraries for Python and presented on a Sony 20" Trinitron monitor positioned 36" from where participants were sitting. Nine different stimuli were used, the products of factorially combining three shades of red (RGB values: $(255, 0, 0), (255, 0, 96), (255, 0, 160)$; determined by pilot testing to produce reasonable ordering of survivor functions) and three different shapes (circles, octagons, and diamonds; Table 3). The target was always a red circle. All stimuli subtended 0.696 degrees of visual angle. The background of the search field was white.

DISTRIBUTION A: Distribution approved for public release.

Table 3: Stimuli used in Experiments 1 and 2

Salience Color	Salience Shape	Name	
A	A	Red Circle	●
A	L	Red Octagon	●
A	H	Red Diamond	◆
L	A	Pink Circle	●
L	L	Pink Octagon	●
L	H	Pink Diamond	◆
H	A	Purple Circle	●
H	L	Purple Octagon	●
H	H	Purple Diamond	◆

Procedure. In each session, participants completed three blocks of trials. The first two blocks one color-only and one shape-only were designed to measure response time distributions when only a single stimulus dimension was manipulated in order to calculate the baseline model for the capacity coefficient. The order of these single-feature blocks was randomized within-session. Each single-feature block consisted of 23 sets of four (2 distractor type: high/low saliency X 2 target presence: present/absent) trials, randomized within-set. In the color-only block, distractors were limited to circles (i.e. target shape; HA and LA stimuli in Table 1). Thus participants were required to find a red circle amongst other circles. In the shape-only block, distractors were limited to red (i.e. target color) shapes (AH and AL stimuli in Table 1). Thus participants were required to find a red circle amongst other red shapes. The third block utilized all eight distractors found in Table 1, and consisted of 37 sets of 16 (8 distractor types X 2 target presence: present/absent) trials, again randomized within-set. This block allowed us to measure response time distributions when both stimulus dimensions were manipulated in order to calculate survivor interaction contrasts. In all three blocks, exactly 24 stimuli were presented on each trial, with an additional distractor replacing the target on target-absent trials. As previously stated, although distractor types varied between trials, within a given trial all distractors were of the same type.

Each trial, regardless of block, followed the same process. At the start of a new trial the words Get ready were displayed in the center of the screen for 1 second, after which the search field would be presented. Stimuli locations were determined on each trial before presentation by randomly placing the target somewhere within the bounds of 100 pixels from the displays edges. Distractor locations were then determined by choosing randomly generated locations within the search field bounds that did not place a distractor within 100 pixels of another stimulus. Trials automatically advanced (i.e. were scratched) if no response was made within 20 seconds. Trials simply advanced to the next trial after a negative (target-absent) response. If a positive (target-present) response was made, the stimuli would be immediately replaced by black outlines of tri-

DISTRIBUTION A: Distribution approved for public release.

angles (chosen because they did not share color or shape information with any of the experimental stimuli), and the mouse, which was not normally visible on screen, would appear in the center of the screen. Such trials would advance only after the participant had clicked on a triangle. This portion of the trial was not timed, although it was used in determining whether a correct response had been made. Response times reported in the results section refer only to the interval between search field onset and time of yes/no response. Responses were scored as correct if and only if: 1) the target was present and the participant responded that the target was present followed by correctly designating the location of the target (by clicking on the appropriate triangle) or 2) the target was absent and the participant responded that the target was not present.

Participants received verbal instructions before the start of the first session as well as on-screen, text-based instructions before each block of every session. The experimenter described the stimuli the participant would encounter in each block, and while they were not explicitly told the ratio of target-present to target-absent trials (1:1), participants were advised that some trials may contain a target and some trials may not. Participants were instructed to determine the presence of the target (red circle) as quickly as possible while maintaining accuracy and to respond yes (target-present) by clicking the left mouse button and to respond no (target-absent) by clicking the right mouse button. Participants were also told that if they responded positively they would then be required to indicate where they found the target by clicking on the triangle that was located where the red circle had been prior to responding. The text-based instructions provided before each block were accompanied by example images (Table 1) of the stimuli and followed by one practice trial for each unique trial condition that would be encountered in the upcoming block (i.e., four practice trials for the first two blocks and sixteen for the third block), in random order. Feedback was provided for these practice trials, including unlimited time to study their mistake: if the response was incorrect on a practice trial, the screen reverted to the original search field, and the target was highlighted if it was present.

After four sessions, each participant had completed a total of 3104 visual search trials, resulting in 92 unique observations of each trial condition in the single-feature blocks and 148 unique observations of each trial condition in the third search block.

Results

Accuracy results are shown in Figure 19. A two-way repeated measures ANOVA comparing accuracy across trial-types and target present/absent indicated a significant interaction between present/absent and trial-type ($F(7, 98) = 26.1, p < 0.001, \eta_G^2 = 0.41$) and main effects of both present/absent ($F(1, 14) = 30.7, p < 0.001, \eta_G^2 = 0.19$ and trial-type ($F(7, 98) = 27.8, p < 0.001, \eta_G^2 = 0.44$). On target absent trials, accuracy was nearly perfect across trial types for all participants, while misses occurred frequently when there was only one distinguishing feature and the salience was low.

Response times showed a similar pattern of significance. The interaction was significant ($F(7, 98) = 30.2, p < 0.001, \eta_G^2 = 0.22$) as were the present/absent

DISTRIBUTION A: Distribution approved for public release.

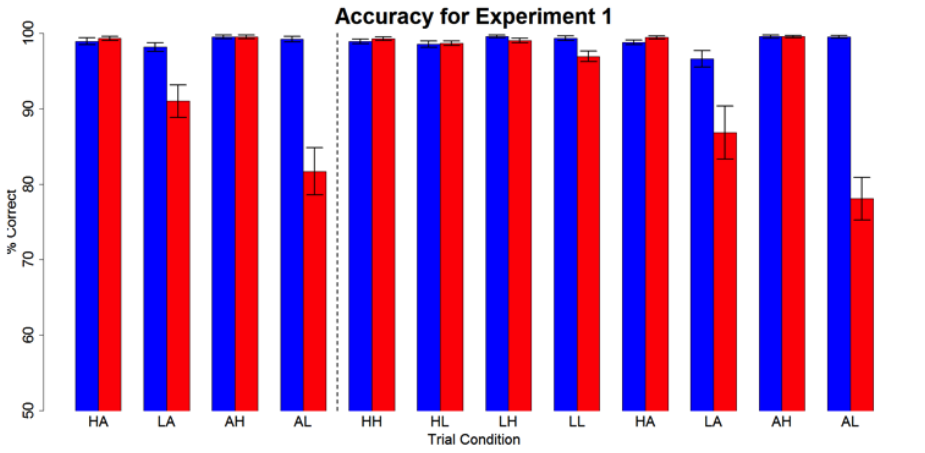

Figure 19: Experiment 1 accuracy. The blue bars on the left in each pair indicate target absent trials and the red bars on the right in each pair indicate target present trials.

main effect ($F(1,14) = 130, p < 0.001, \eta_G^2 = 0.23$) and trial-type main effect ($F(7,98) = 88.7, p < 0.001, \eta_G^2 = 0.76$).

SIC results for target absent and present trials are shown in Figure 20. In the both the target absent and target present trials, none of the 15 participants showed evidence of a violation of selective influence based on the series of pairwise KS tests. In the target absent trials, all 15 participants had significantly positive MIC and SIC values and 7 participants also had a significantly negative SIC. In the target present trials, all 15 participants had significantly positive MIC and SIC values and none had significantly negative SIC values.

Our first experiment successfully characterized the display-level integration of color and shape information in the context of visual search with homogeneous distractors. However, because the design of Experiment 1 allowed for accurate task completion without forcing exhaustive processing of both stimulus features on every trial, the full extent to which color processing and shape processing interact remains uncertain. Experiment 2 sought to definitively answer this question by altering the previous paradigm to instead use heterogeneous distractors. That is, we continued to manipulate the prevalent distractor on each trial, but we also presented every possible remaining distractor as well. Because all distractor types were present on every trial, participants were forced to process both color and shape information if they were to maintain a high level of accuracy throughout the experiment.

The study was conducted at Wright State University and approved by its Institutional Review Board. Pilot testing revealed that participants were generally slower to respond with the new paradigm than they were in Experiment 1. Therefore, in order to obtain approximately the same number of unique ob-

DISTRIBUTION A: Distribution approved for public release.

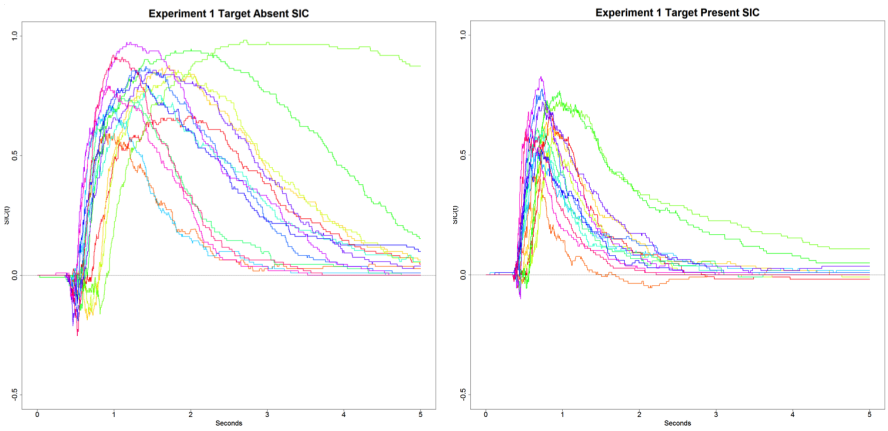

Figure 20: Survivor interaction contrast functions for target absent (left) and target present (right) response times in Experiment 1. Each line indicates an individual participant.

servations in each trial condition that we gathered in the previous experiment without increasing the duration of each experimental session, we administered Experiment 2 over five one-hour sessions. As before we did our best to ensure all five sessions took place within the period of one week and no more than two weeks.

Participants. All participants gave written informed consent before beginning the study and were compensated $8 per one-hour session. Participants who completed all five sessions were awarded an additional $2 per session bonus (i.e. $50 maximum compensation). Twenty members of the Wright State University community were recruited to participate, sixteen of whom completed all five experimental sessions. Data from one of these sixteen participants was discarded for exceptionally poor accuracy, leaving data from fifteen participants available for further analysis.

Materials. Materials used in Experiment 2 were identical to those used in Experiment 1 (Table 1).

Procedure. Instructions and procedure for Experiment 2 were identical to the previous experiment except for the previously mentioned additional session and the change to heterogeneous distractors. As in Experiment 1, the search field of every trial contained 24 stimuli; however, every distractor type (available in the respective block) was present on each trial, with one type of distractor always more common than the others.

As before, participants completed a color-only and a shape-only block, randomized within-session, in order to calculate the baseline model for capacity analysis. Each single-feature block consisted of 16 sets of four (2 prevalent distractor type: high/low saliency $\times$ 2 target presence: present/absent) trials, randomized within-set. In the single-feature search blocks, where there were

DISTRIBUTION A: Distribution approved for public release.

only two unique distractor conditions, the prevalent distractor type made up 15 of the distractors while the alternative type made up 8 of the distractors (e.g. a shape-only trial might contain 1 red circle, 15 red octagons, and 8 red diamonds, reflecting the target-present, absent-high saliency trial condition).

The third block of Experiment 2 was a conjunctive search task, where participants were required to distinguish both target features from distractors because stimuli that shared a feature with the target were present on every trial. Participants completed 30 sets of 16 (8 prevalent distractor types $\times$ 2 target presence: present/absent) trials, again randomized within-set. On a given trial, the prevalent distractor type accounted for 9 of the distractors while the remaining types each made up 2 of the distractors (e.g. the target-present, high-high saliency trial condition contained 1 red circle, 9 "purple" diamonds, 2 "purple" octagons, 2 "purple" circles, 2 "pink" diamonds, 2 "pink" octagons, 2 "pink" circles, 2 red diamonds, and 2 red octagons). On target-absent trials, an additional distractor of the prevalent type for the trial condition replaced the target.

Each block was preceded by instructions and practice trials, as in Experiment 1; however, because the goal of the practice trials was to familiarize the participant with the response procedure rather than the stimuli themselves, only two practice trials (one target-present and one target-absent, order randomly determined) were used. On these trials the proportion of distractor types were equated such that twelve each of the two types available in the singe-feature blocks were presented for those blocks and three each of the eight distractor types in the conjunctive search block were presented. To preserve this balance and prevent bias, the target on target-present practice trials became the 25th stimulus. These practice trials were the only time 24 stimuli were not presented.

After five sessions, participants had completed a total of 3040 visual search trials each, resulting in 80 unique observations of each trial condition in the single-feature blocks and 150 unique observations of each trial condition in the conjunctive search blocks.

Results

Accuracy results are shown in Figure 21 and were similar to Experiment 1, although the effect sizes were reduced. A two-way repeated measures ANOVA comparing accuracy across trial-types and target present/absent indicated a significant interaction ($F(7,98) = 12.5, p < 0.001, \eta_G^2 = 0.043$) and main effect of present/absent ($F(1,14) = 7.40, p = 0.017, \eta_G^2 = 0.16$) and trial-type ($F(7,98) = 3.42), p = 0.003, \eta_G^2 = 0.02$). As in Experiment 1, accuracy was nearly perfect across trial types for all participants ontarget absent trials, while misses occurred frequently when there was only one distinguishing feature and the salience was low.

Response time results were also similar to Experiment 1 as were the associated effect sizes. The interaction was significant ($F(7,98) = 43.3, p < 0.001, \eta_G^2 = 0.13$) as were the main effects of present/absent ($F(1,14) = 75, p < 0.001, \eta_G^2 = 0.55$) and trial-type ($F(7,98) = 142, p < 0.001, \eta_G^2 = 0.45$).

SIC results for target absent and present trials are shown in Figure 20. In the both the target absent trials, only 1 of the 15 participants showed evidence of a violation of selective influence based on the series of pair-wise KS tests but

DISTRIBUTION A: Distribution approved for public release.

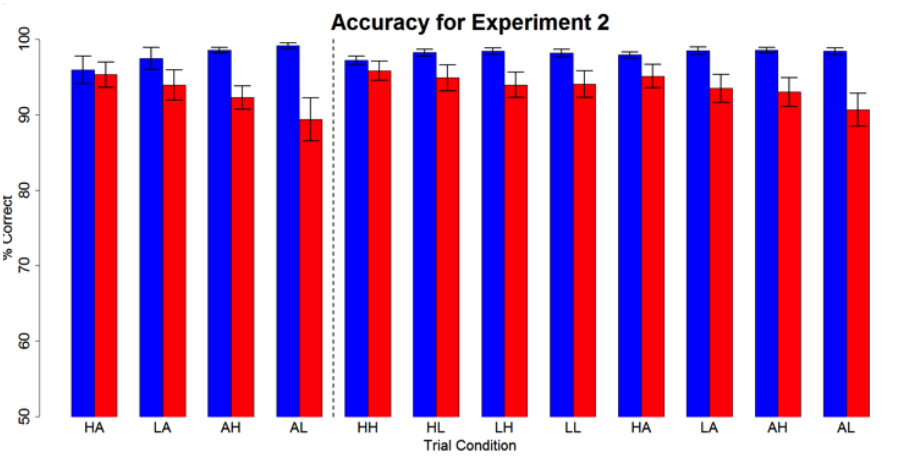

Figure 21: Experiment 2 accuracy. The blue bars on the left in each pair indicate target absent trials and the red bars on the right in each pair indicate target present trials.

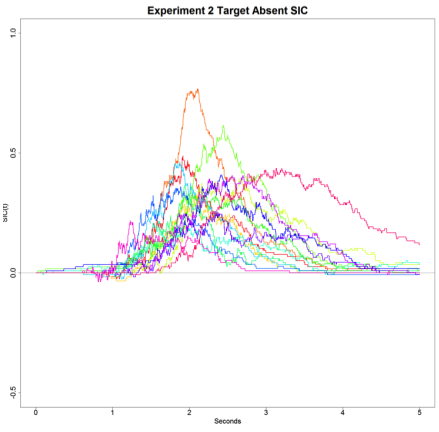

Figure 22: SIC functions for target absent response times in Experiment 2. Each line indicates an individual participant. SICs from the target present trials are not shown because all 15 showed violations of selective influence.

DISTRIBUTION A: Distribution approved for public release.

all 15 subjects showed evidence of a violation in the target present trials. In the target absent trials, all 14 participants who did not violate the selective influence assumption had significantly positive MIC and SIC values. None of those 14 participants had significantly negative SIC values.

Based on these experiments, color and shape processing during single-feature and conjunctive visual search is parallel and highly facilitatory, potentially even coactive.

5 The Perception of Fused Multispectral Imagery[8]

When information across two sensors is for the most part redundant, multisensor fusion hinders performance, regardless of whether they are presented side-by-side or fused into a single composite image. An observer may instead benefit from the use of one single-sensor image that provides the requisite information to make a accurate, quick decision. If both sensors images are displayed, presenting the images side-by-side leads to less inefficient performance than algorithmically fused images. With the particular imagery we used, it is clear that the inefficient performance with the side-by-side imagery is not due to serial processing or to waiting to complete processing of both sources. Instead, the limitation is more likely due to attentional or other intrinsic limitations.

In general, future research on image fusion should include more sophisticated baselines than just performance with single-sensor imagery. Model-based empirical design approaches, particularly SFT, illuminate differences in the efficiency with which observers combine information across sensors. Furthermore, SFT can be used to determine whether inefficiencies are due to strategic factors, such as using sensor images in serial checking both images regardless of whether one is sufficient, or due to other intrinsic limitations.

Information from non-visible parts of the electromagnetic spectrum is beneficial for determining different types of environmental information in many operational settings (Hall & Llinas, 1997). For example, long-wave infrared (LWIR) emissions are useful for detecting heat information (e.g., occluded heat producing objects such as a person behind a bush), and short-wave infrared (SWIR; e.g., night vision) can pick up detail in conditions with low illumination. Together, infrared and visible sensors may supply the operator with complementary information and aid in a task such as determining a target's (e.g., person) location relative to an object in the scene (Toet, Ljspeert, Waxman, & Aguilar, 1997).

There are several alternative ways to present an observer with multiple sensor images simultaneously. A common family of approaches, which we refer to as *algorithmic* fusion, is to combine relevant information from two sensor images into one composite image (Burt & Kolczynski, 1993). Alternatively, information from each sensor could be displayed in two separate images. Presenting all

[8]Content in this section is from Fox and Houpt (2016) and Zhang and Houpt (2016).

DISTRIBUTION A: Distribution approved for public release.

available information moves the choice of relevant information to the operator rather than relying on an algorithm to detect useful sensor information.

Algorithmic fusion has been the focus of much of the research on presenting multi-spectral information. This is due to two potential benefits of the technique: 1) algorithmic fusion restricts number of sources of visual information to which the operator must attend; and 2) the resultant image may possess emergent features not found in either single image alone (Krebs & Sinai, 2002). A potential downside to algorithmic fusion is that some information from the individual sensors must be filtered out in the process of creating a single image (Hall & Steinberg, 2000). There are many options for algorithmic fusion, and the choice of algorithm does offer some freedom in determining what information is lost, but information is necessarily lost.

In some domains, giving complete information to an operator, particularly expert operators, leads to advantages (cf. Klein, Moon, & Hoffman, 2006). In the image fusion literature, the process of an operator using information from multiple separate images for a task is often referred to as "cognitive fusion" (cf. Blasch & Plano, 2005) because any potential integration of the two images must take place cognitively. Cognitive fusion is a moniker we will adopt for the rest of this paper. Note that cognitive fusion refers to performance using separate images, not necessarily a particular form of cognitive or perceptual process.

In this paper, we suggest the use of a cognitive-theory-driven approach based on performance, systems factorial technology (SFT), for evaluating image fusion approaches, particularly for comparing algorithmic to cognitive fusion. This approach allows for both more theoretically meaningful measures than raw accuracy or response time (RT), and for insight into the particular aspects of the cognitive process that may have led to better or worse performance. We will begin by briefly reviewing the existing approaches to evaluating image fusion. Next, we review SFT, then apply the methodology to compare algorithmic (in this case Laplacian pyramid fusion, which we describe below) fusion to cognitive fusion (side-by-side image presentation).

Image fusion is mostly studied within the field of computer vision, hence the vast majority of the metrics of fusion quality are based on computational principles. One of the more common measures is of the preservation of edge information (either at the individual pixel level Xydeas & Petrović, 2000; the local, 8×8 pixel grid level Piella & Heijmans, 2003; or the global image level Petrović & Xydeas, 2004; Qu, Zhang, & Yan, 2002). These image-level metrics are valuable in that they provide an objective assessment of the amount and quality of information from each single-sensor that is represented in the composite image for minimal cost. Two major deficits of limiting assessment to image quality metrics is that they do not account for task relevant information and are not always predictive of human performance (Smeelen, Schwering, Toet, & Loog, 2014).

To address the shortcomings of computer based image quality metrics, subjective user experience questionnaires (asking for example, overall reported image preference, comfort, etc.) are used (Krishnamoorthy & Soman, 2010; Petrović, 2007). This approach offers a partial solution, but subjective qual-

DISTRIBUTION A: Distribution approved for public release.

ity assessments can also fail to predict variation in performance. Furthermore, when they are used, user experience assessments are only used for outcome assessment and not to directly inform the design process (Toet et al., 2010b). Hence, while subjective quality of a display yields some benefits, to gain understanding of what design aspects leads to better decision-making and human performance and inform the design of new fusion approaches, it is important to directly measure human performance on a specific task(cf. Blum, 2006; Dixon et al., 2006; Dong, Zhuang, Huang, & Fu, 2009).

Despite being a relatively limited literature, human performance with fused imagery has been used with a range of basic visual tasks including detection (Krebs et al., 1999), discrimination (e.g., global scene is upright or vertically inverted; Krebs & Sinai, 2002; Toet et al., 1997), recognition (Ryan & Tinkler, 1995; Sinai, McCarley, & Krebs, 1999; Toet & Franken, 2003), and visual search (Neriani, Pinkus, & Dommett, 2008). This research has been conducted in contexts including aviation (Ryan & Tinkler, 1995; Steele & Perconti, 1997) and surveillance (Neriani et al., 2008; Toet & Franken, 2003; Toet et al., 1997). Among these applications, there is a wide range of reported results and overall conclusions. Such discrepancies are potentially due to methodological variation (Ahumada & Krebs, 2000; Essock, Sinai, McCarley, Krebs, & DeFord, 1999; Steele & Perconti, 1997), differences in task descriptions (Krebs & Sinai, 2002; McCarley & Krebs, 2000), and variation in fusion algorithms or sensor combinations (McCarley & Krebs, 2000; Neriani et al., 2008). Additional manipulations often cited in the literature are task type and difficulty, image scene, sensors, and fusion algorithms (Krebs & Sinai, 2002; McCarley & Krebs, 2000). Thus far there is no standard way to compare across manipulations that controls for the amount and type of information provided by each component image.

In many of these studies, performance with composite images was compared to performance with an individual sensor (e.g., long-wave infrared + visible compared to visible-alone). Unfortunately, this comparison confounds whether image fusion enhances performance because of the fusion method implemented or simply because it supplies more information to the observer. We are concerned with answering the question of whether the observer is processing each sensor image as efficiently in a multi-sensor context as when presented in isolation. To effectively answer this question we must compare performance with multiple sensors to a prediction of how well they should perform given their performance with each individual sensor image.

When an observer is provided two sensor images, regardless of the display type, they have redundant information to inform them of the correct decision, thereby suggesting an overall faster response. Although it may seem intuitive to equate a performance gain with redundant signals with facilitatory processing, parallel processes with no facilitation can predict significant redundancy gains (Duncan, 1980; Kahneman, 1973; Miller, 1982; Raab, 1962; Townsend & Wenger, 2004b). Furthermore, performance decrements may still be observed relative to single-source imagery due to our perceptual system dealing with multiple pieces of information (cf. Townsend & Ashby, 1983b; Townsend & Wenger, 2004b). Thus, it is important to use an appropriate baseline for assessing the gain (or

DISTRIBUTION A: Distribution approved for public release.

loss) due to an added signal. The capacity coefficient, a measure from SFT that we describe in detail in the next section, addresses this issue because it uses individual source performance to predict what performance would be in a multi-signal context under a baseline model assumption.

By using SFT, we go beyond the simple better/worse distinctions that are possible with the previously applied metrics. SFT allows us to examine the reason for observed performance differences including: differential effects of increasing the amount of available information (i.e., processing efficiency); facilitation or inhibition between the perception of each source of information; whether processing one image source is sufficient or both sources must be processed; and the temporal organization of the perception (i.e., serial versus parallel).

The use of SFT allows us to examine the underlying processes to help explain why we may see performance benefits of a particular operator display. Each variation in processing structure may inform the cause for a particular pattern of performance. If participants are presented with task relevant yet redundant information across sensors they may adopt a processing strategy in which information from only one sensor is used to make the decision (i.e., "OR" processing or first-terminating). OR processing may combine with either a parallel- or serial-processing structure: either information from both sensors is processed simultaneously but only the fastest to finish is used to make the discrimination (parallel-OR) or information from one sensor is processed and is used for the decision while the alternative sensor is not processed (serial-OR). Alternatively, individual sensor images may each contribute unique, complementary information forcing participants to process both sensors entirely to make a correct decision ("AND" processing). AND processing may also combine with either a parallel- or serial-processing structure: both sensors are processed simultaneously and the slowest to finish is used to make the discrimination (parallel-AND) or both sensors are fully processed, first one, then the other (serial-AND). Fusion also allows for a single percept in which all information is processed in parallel and is pooled to make a decision (coactive processing).

Here we discuss what particular processing mechanisms suggest on a more conceptual level about visual cognition for each presentation type: algorithmic and cognitive fusion.

For algorithmically fused images, standard serial and parallel architectures may be possible, although are a priori unlikely. An interpretation of such finding would be that participants can selectively attend to one particular spatial frequency information based on the distinctive features to complete the task (Morrison & Schyns, 2001). Alternatively, if observers are unable to selectively extract information from each perceptual dimension, as indicated by McCarley and Krebs (2006), then a coactive or interactive parallel process is more likely (Eidels, Houpt, Pei, Altieri, & Townsend, 2011b, cf.). For algorithm-fused imagery we hypothesize: 1) individuals' efficiency will be at least as high as respective UCIP predictions (i.e., unlimited capacity) across all discrimination stimuli, and 2) individuals' will use a highly interactive, parallel mechanisms for processing the multi-sensor information.

When images are presented beside one another (i.e., cognitive fusion) peo-

DISTRIBUTION A: Distribution approved for public release.

ple may process each sensor image in series or in parallel. If processing both images requires visual attention shifts between the two images, then it may be more likely that the images are processed in series. This mechanism limits performance by the constraints of mental integration across several samples of information (Irwin, 1991; Rayner, McConkie, & Zola, 1980). However, serial processes can lead to efficient processing if information from only one image is sufficient for adequate judgments and the additional image is redundant and potentially unnecessary (Neriani et al., 2008).

Alternatively, people may process and potentially integrate the two images in parallel, leaving the opportunity for facilitation in judgment performance due to pictorial redundancy speed-ups (Pollatsek, Rayner, & Collins, 1984), which would imply facilitatory parallel or coactive processing. In contrast, if processing the information across two images is a larger drain on attentional resources, degrading performance with each image (Rousselet, Fabre-Thorpe, & Thorpe, 2002; Scharff, Palmer, & Moore, 2011), inhibitory parallel processing would be observed. Our hypothesis for cognitive fusion focus on predicting a processing strategy that yields: 1) performance no worse than algorithmic fusion. Therefore, individuals' efficiency will be at least as high as respective UCIP predictions (i.e., unlimited capacity) and across all discrimination stimuli, and 2) individuals' will use efficient parallel mechanisms for processing the multi-sensor information.

The cognitive processes involved with utilizing information from multiple sensors may vary from the processing of one sensor image. A cognitively-motivated baseline model can encode a specific set of processes so that systematic deviations from the baseline will give evidence for how the processes have changed. Furthermore, using a standardized method to assess deviations of actual performance from predicted performance given the individual parts yields a flexible approach to make comparisons of human processes across several experimental manipulations such as alternative sensors, stimuli, and fusion methods.

5.1 Weapon/Tool Discrimination

The goal of the first experiment was to measure the capacity coefficient with both side-by-side imagery and algorithmically fused imagery.

Ten observers whom gave informed consent participated in the study (male = 6, average age=23.8). All had normal or corrected-to-normal visual acuity, and normal color vision. After finishing the study, each participant received 10 dollars as compensation. Stimuli. Stimuli were ten images taken of a female holding either a gun or a tool using both a long-wave infrared (LWIR) sensor and a standard visible-spectrum sensitive camera. Both visible and LWIR images were mapped to grey scale for presentation to the observers. Fused images were created using simultaneously captured LWIR and visible images that were fused image using a Laplacian pyramid algorithm (see B. Yang, Jing, & Zhao, 2010, for a review). All images were 256×256 pixels in size. Stimuli were presented in the center of a 19 monitor with resolution of 1280 1024 pixels and a refresh

DISTRIBUTION A: Distribution approved for public release.

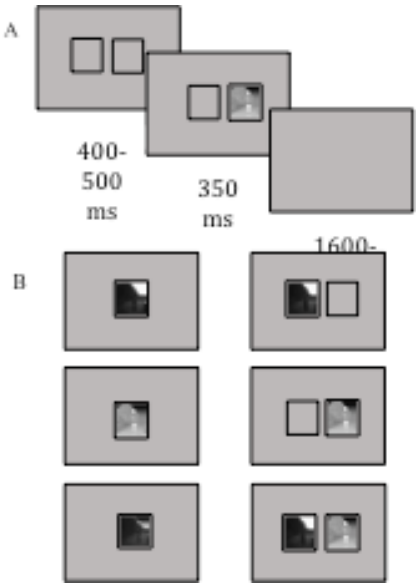

Figure 23: A: Basic trial structure. B: Types of stimulus organization used in the experiment, either central/fused (left) or side-by-side (right).

rate of 85 Hz.

Figure23 gives an overview of the trial procedure. The observers task was to click mouse to indicate whether the person in the image was holding a gun or tool. Each trial began with a square indicating where images might appear that lasted a uniformly random duration between 400ms and 500. Next, the stimuli appeared for 350ms followed by a blank screen while waiting for the subjects response. There were six different blocks totaling 1200 trials. The whole session lasted approximately one hour.

Accuracy and correct response times were analyzed with a repeated measure ANOVA using the ez package (Lawrence, 2012) in R (R Development Core Team, 2011). One ANOVA was run to assess the difference between a single sensor image in isolation and two sources presented together (averaged across both fused a side-by-side presentation). A second ANOVA was applied the subset of the data for which both sensor were presented to assess the difference between fused and side-by-side images. The third ANOVA checked all the other factors that would lead to differences for non-fused blocks.

The difference between single-source images ($M = 515.83, SD = 176.33$; $M = 0.93, SD = 0.26$) and two-source images ($M = 530.92, SD = 183.89; M = 0.91, SD = 0.29$) was neither significant for response times ($F(1, 9) = 3.93, p = 0.08$) nor accuracy ($F(1, 9) = 1.48, p = 0.26$).

The difference between performance with cognitive fusion (M = 542.19, SD = 179.78; M = 0.94, SD = 0.24) and algorithmic fusion (M = 518.87, SD = 187.52; M = 0.88, SD = 0.33) was not significant for either response times (F

DISTRIBUTION A: Distribution approved for public release.

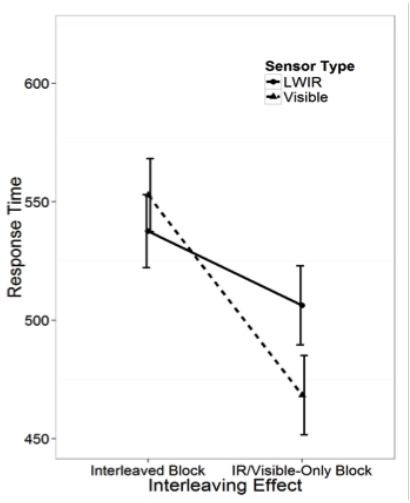

Figure 24: Response time results for interleaved trials versus single-sensor blocks crossed with sensor type. Error bars indicate confidence intervals calculated based on Jarmasz and Hollands (2009).

(1, 9) = 1.21, p = 0.3) or accuracy (F (1, 9) = 2.92, p =0.12).

An ANOVA tested sensor type of images, location presented, and interleaving effect in non-fused blocks. Only a main effect of interleaving effect was significant for accuracy ($F(1,9) = 19.60, p < 0.01, \eta_G^2 = 0.06$); for response times, when images were presented in the center of the screen, subjects responded faster ($M = 497.98, SD = 169.40$) than when images were presented side-by-side ($M = 534.05, SD = 181.33$; $F(1,9) = 19.68, p < 0.01, \eta_G^2 = 0.10$). Performance was better in the no interleaving (only LWIR or visible images) blocks ($M = 487.39, SD = 159.50$) than interleaved blocks ($M = 545.20, SD = 487.39; F(1,9) = 40.68, p < 0.01, \eta_G^2 = 0.10$). The main effect of sensor type, LWIR ($M = 509.90, SD = 180.47$) or visible ($M = 521.71, SD = 171.93$), was not significant ($F(1,9) = 1.63, p = 0.23$). But there was a significant (cross-over) interaction with the interleaving effect (Figure 24, $F(1,9) = 24.53, p < 0.01, \eta_G^2 = 0.02$). And interleaving effect also interacted with the location presented (Figure 25; $F(1,9) = 42.87, p < 0.01, \eta_G^2 = 0.04$).

From Figure 24 and Figure 25, we could see that when LWIR and visible images were interleaved in presentation, subjects couldnt predict the next images type, thus the advantage of visible image ($M = 468.40, SD = 152.54$) over LWIR image ($M = 506.28, SD = 164.00$) disappeared. Also, in the interleaved block, subjects needed to switch attention between the left side and right side, so people responded slower ($M = 583.12, SD = 194.28$) compared to that when they only need to focus on the center of screen ($M = 508.18, SD = 173.28; F(1,9) = 35.83, p < 0.01, \eta_G^2 = 0.80$).

Capacity analyses were only applied to observers who had at least 90% ac-

DISTRIBUTION A: Distribution approved for public release.

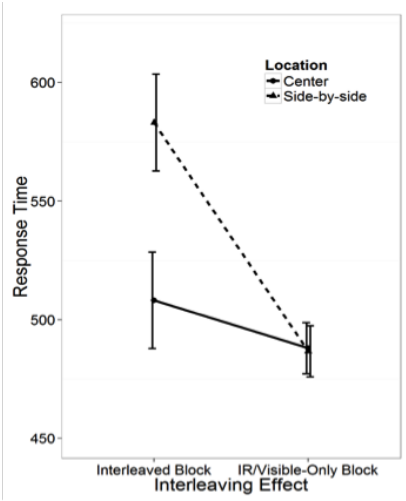

Figure 25: Response time results for interleaved trials versus single-sensor blocks crossed with stimulus organization. Error bars indicate confidence intervals calculated based on Jarmasz and Hollands (2009).

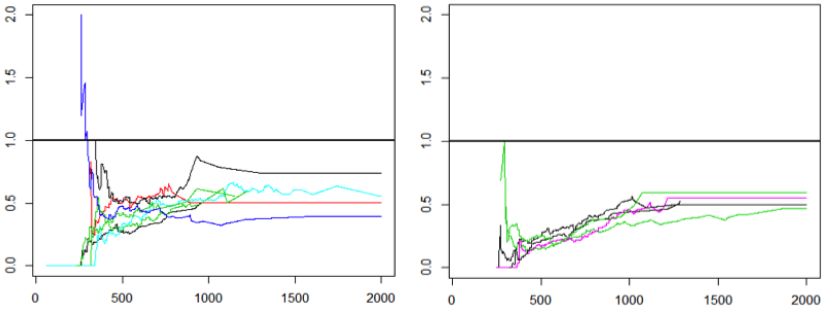

Figure 26: Capacity coefficient results for algorithmically fused images (left) and side-by-side images, i.e., cognitive fusion (right).

DISTRIBUTION A: Distribution approved for public release.

curacy in all conditions within the algorithmic blocks (7/10 participants) or cognitive-fusion blocks (5/10 participants). Figure 26 depicts those participants capacity coefficients. Individual level capacity analysis indicated seven participants were limited capacity (z-scores between -11.16 and -10.27) with cognitive fusion and five were limited capacity with algorithmic fusion (z-scores between -10.98 and -6.13). Group level t-tests indicated both fusion methods led to limited capacity (Algorithmic: $t(6) = -12.64$; Cognitive: $t(4) = -73.36$). Cognitive fusion was statistically more limited compared to algorithmic fusion ($t(6.62) = 4.02, p < 0.01, d = 2.16$).

The current results are consistent with previous research in our lab (Fox, 2015). Both algorithmic fusion and cognitive fusion are capacity limited, and the efficiency of cognitive fusion is at least as high as algorithmic fusion. The process of switching attention between locations increased reaction time and reduced the advantage of visible images over IR images in our task. The side-by-side presentation method increased reaction times but resulted in equal performance compared to the images developed by algorithmic fusion in the center of screen.

There are various potential explanations for the limited capacity performance that we observed. For the side-by-side imagery, observers may have only used one source and hence lost out on the redundancy gain predicted by independent parallel processing (cf. Raab, 1962). Indeed we believe the lack of a redundancy gain is the reason for limited capacity performance with fused images; because the information in both sensor images is essentially the same, the fused image is not necessarily more informative than either individual source image. If observers were using a strategy of focusing on only one sensor image, then they should have a flat SIC and 0 MIC when examined with the factorial salience conditions. Alternatively, observers could be using both sensor images but, due to limited resources (e.g., attention, foveation), be processing each source slower when they are together. In this case, observers would have a positive SIC and MIC. To discriminate between these two possible explanations of performance with the side-by-side imagery, we ran a follow-up experiment with the factorial manipulations necessary to calculate the MIC and SIC.

Because the manipulation of each sensor image is no longer selective after algorithmic fusion, we were not able to apply the SIC and MIC analysis to the algorithmically fused imagery.

The goal of Experiment 2 was to measure the SIC and MIC from participants when they were making discrimination judgments with side-by-side imagery.

Methods

Observer. Ten new observers who gave informed consent participated in the study (male = 3, average age=23.7). All had normal or corrected-to-normal visual acuity, and normal color vision. After finishing the study, each participant received 40 dollars as compensation.

Stimuli. The base images were the same as images in Experiment 1 (although the algorithmically fused images were not used). For slow processing trials, the base images were displayed with zero mean Gaussian luminance noise added. The variance of the noise was chosen individually as described in the next subsection.

DISTRIBUTION A: Distribution approved for public release.

Table 4: SIC Analysis for Weapon/Tool Experiment 2

Subject	Tool			Weapon		
	SIC	MIC	Strategy	SIC	MIC	Strategy
1	+	+	P-A	+	+	P-A
2	NA	NA	NA	+	+	P-A
3	-	=	P-O	=	=	S-O
4	NA	NA	NA	+	=	P-A
5	+	+	P-A	+	+	P-A
6	-	=	P-O	+	+	P-A
7	-	=	P-O	+	+	P-A
8	+	+	P-A	+	+	P-A
9	+	+	P-A	+	+	P-A
10	+	+	P-A	-	-	P-O

Note: For strategy, P-A means parallel-AND, P-O means parallel-OR, and S-O means serial-OR.

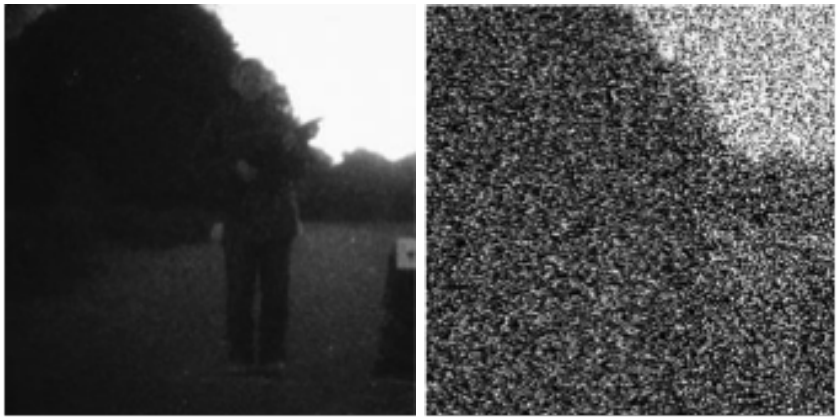

Figure 27: Example image from a visible sensor without noise (left) and with noise (right). Raw images were used for the high-salience stimuli and images with added noise were used for the low salience stimuli.

DISTRIBUTION A: Distribution approved for public release.

Procedure. Each trial was the same as Experiment 1. For each trial, images appeared on left side, right side or both. There were two blocks in each session; the first block used Psi adaptive psychophysical method (Kontsevich & Tyler, 1999) manipulating noise added to the image that gets 90% accuracy of performance. The image with noise was referred as slow processing condition and the image without noise was referred as fast processing condition (Figure 27). In the second block, there were eight different combinations in condition of slow processing, fast processing, and absent (no image revealed) controls. Each session consisted 1320 trials that last about one hour, and it repeated for four days.

Results

We separated gun from tool in stimuli for SIC analysis. Two subjects were excluded for tool SIC analysis because they failed to pass selective influence. Five of eight subjects SIC were statistically significant positive; three of them were statistically significant negative according to the HouptTownsend SIC statistic (Houpt & Townsend, 2010a). When gun was presented, eight of ten subjects SIC were statistically significant positive. Neither the negative part of the SIC nor the positive part was significantly different from zero for one subject, another subjects SIC was significantly negative.

Discussion

The results for the experiment indicated only four observers employed same parallel-AND processing strategy for identifying gun and tool. For the four subjects whose both tool and gun SIC data were interpretable: two subjects used parallel-OR strategy for identifying tool and parallel-AND for identifying gun, one subject used parallel-OR for identifying tool and serial-OR for identifying gun, one subject used parallel-AND strategy for identifying tool and parallel-OR for identifying gun. Based on the simulation studies reported in (Houpt & Townsend, 2010a) and the estimated SICs, it is unlikely that those four observers used the same strategy for both tool and gun.

GENERAL DISCUSSION

Our first experiment indicated that the information from additional source does not gain advantage in information processing as expected. Our second pilot experiment tried to explore the reason for limited capacity in cognitive fusion in Experiment 1. It suggested searching strategy might lead to inefficiently processing when images from different sensors were presented side-by-side. When identifying gun, the observers may be limited capacity because they are waiting to process both images rather than responding as soon as they have identified gun in either image. When identifying tool, the observer may be limited capacity due to serial-OR processing (i.e., only using one image source) or they may have been parallel-OR (and the SIC was not significant due to lack of power) and the limitation may be due to limited perceptual resources. Both algorithmic and cognitive fusion methods are limited in capacity, and cognitive fusion method is more limited in current the weapon/no-weapon task. Thus we suggest that cognitive fusion method such as side-by-side presentation should be considered as an alternative way in image fusion design. Side-by-side presentation does not necessarily lead to worse performance but dependent upon the task (Dixon et al., 2007). Operators should have a choice for different fusion

DISTRIBUTION A: Distribution approved for public release.

methods dependent upon the task being undertaken and situation they are in.

5.2 Human Orientation Discrimination

There was substantial overlap in the methods across the next two experiments. First we outline the common methods below then give experiment specific details in their respective sections.

The trials for the SIC were collected in a separate block from those blocks that were included for estimating the capacity coefficient. This allowed us balance the number of trials in such a way as to not bias responses to one source based on the other source (conditioned on the stimulus) in accordance with the constraints outlined in Houpt et al. (2012) following Mordkoff and Yantis (1991b).

To estimate the capacity coefficient, we need RTs from trials in which participants can respond to both visible and LWIR images (i.e., either algorithmically or cognitively fused imagery) as well as trials in which they are only focused on a single source (i.e., visible only or LWIR only). To get the best estimate of what UCIP performance would be, trial type was blocked. Hence, each participant had a block that was entirely dedicated to visible imagery, a separate block dedicated to LWIR imagery and a block dedicated to fused imagery.

For capacity analyses, we used the imagery without any added noise, which corresponded to the high salience (H) conditions in the SIC analysis (outlined in the "Survivor Interaction Contrast" section above). Recall, the order of the elements in the subscript indicates the source of information, with the first subscript indicating the LWIR information and the second indicating the visible information. Hence, we denote the visible only trials with the subscript $\emptyset H$, the LWIR trials with $H\emptyset$ and the fusion trials with HH.

To estimate the SIC, we need RTs from each factorial combination of source image salience (i.e., with or without added noise). To appropriately interpret the SIC, the salience manipulations must satisfy the assumption of selective influence: the presence or absence of noise added to a source image (e.g., LWIR) should affect the perception of that source but not the other source (e.g., visible).

All participants were recruited from the Wright State University community and gave informed consent consistent with standard ethical guidelines. These experiments were approved by the Wright State Institutional Review Board.

All participants self-reported right-handedness, normal or corrected to normal visual acuity, normal color vision, and no difficulties reading English.

Stimuli were presented using PsychoPy (Peirce, 2009) on a 20-inch Sony Trinitron monitor. Participants sat at a table 75cm from the monitor. Responses were made using a right or left click on a two-button mouse.

The base images were collected using the TRICLOBS 3-band night vision system consisting of two digital image intensifiers (Photonis ICU's) and an uncooled long-wave infrared microbolometer (XenICS Gobi 384) constructed by TNO Defense located in Soesterberg, Netherlands (Toet, 2013). The sensor suite registers visual (400-700 nm), near infrared (700-1000 nm) and long-wave infrared (8000-14000 nm) bands of the electromagnetic spectrum. For this study,

DISTRIBUTION A: Distribution approved for public release.

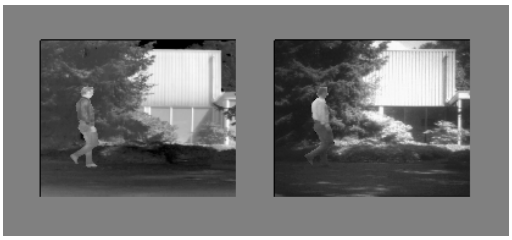

Figure 28: Example of a cognitive fusion presentation of LWIR (always left) and visible (always right). The participants were asked to discriminate whether the person was facing to their left or right. The two images were centered and presented within 6.39° of visual angle on a mid-gray background.

we used imagery from the visible and LWIR sensors as they represent the most distinct ranges of the EM spectrum in this image set and hence potentially carry the most distinctive information.

The optical axes of the three cameras were aligned to minimize the need for registering the images from each sensor post collection, although further registration was done with software developed by Toet and colleagues (Toet & Hogervorst, 2009). Additional image registration was conducted at the Air Force Research Laboratory. Images were approved for public release (Distribution A: Approved for public release; distribution unlimited. 88ABW Cleared 11/18/2014; 88ABW-2014-5325).

We used the Laplacian Pyramid Transform (LPT; Burt & Adelson, 1983) to combine the visible and LWIR information into one composite image. Subjective and image quality assessments support the use of LPT (Petrović, 2007). The LPT is a pixel-level, pyramid-based algorithm meaning we utilized six band filters to pass across both sensor images resulting in a series of image components at different resolution qualities. The component images were averaged together across sensors at each band-pass level and combined using a Laplacian transform. The resultant image was a single composite image containing information from both individual sensors (see Figure 29 for an example).

Note that, as evident in Figures 29 and 30 the combination of the LWIR and visible image using the algorithm does not necessarily enhance an image and may actually degrade the quality of the composite representation. Often times, added image enhancement techniques are used to provide benefits above raw algorithmic fusion. In our study, we use only the existing algorithm supported in the literature to simulate a more real-world environment where the particular task information, and in turn how to further enhance this information, is unknown before displaying the composite algorithmic image.

To compute the SIC, we needed to selectively speed up and slow down the processing of information for both LWIR and visible images while allowing participants to maintain high accuracy. To reduce the image salience, and hence slow processing, we added zero mean luminance noise to the image. An example of a LWIR image and a visible image with white noise is shown in Figure 29.

DISTRIBUTION A: Distribution approved for public release.

To determine the largest amount of noise that we could add without causing accuracy to drop below 90%, we used the QUEST psychometric method (A. Watson & Pelli, 1983). Each SIC session began with 120 trials for each single source image type with varying levels of noise determined by the QUEST adaptive procedure. This allowed us to set individualized salience levels that were specific to each day. Thresholds were estimated each day to account for possible learning and other sources of variation across days. Whether visible only or LWIR only was first was randomly chosen across days and participants.

For computing the SIC, original stimuli (high salience or H) and stimuli with noise (low salience or L) were factorially combined to speed up and slow down the processing of each single-sensor. Factorially combining the images led to four unique multi-sensor combinations: High–LWIR + High–visible, High–LWIR + Low–visible, Low–LWIR + High–visible, and Low–LWIR + Low–visible. For algorithmically fused trials (Experiment 1 only), the stimulus noise was added before fusing the two images together.

Each experiment consisted of 10 days of 1-hour sessions. All participants were compensated $8 per session with a $2 per session completion bonus: $8 + $2 bonus x 10 days = $100 in total for each experiment.

The algorithmically fused images were always presented in the center of the screen within 2.86° of visual angle. For cognitive fusion, both single-sensor images were simultaneously presented 0.67° apart (inner-edge to inner-edge) within 6.39° of visual angle on the screen and directly to the left and right of center screen (cf. Figure 28).

At the beginning of each trial, either a single localization box was shown in the center (algorithmic fusion blocks) or two boxes were presented side by side (cognitive fusion blocks). Localization boxes were always presented for a random interval of time between 400 and 500 msec followed by the stimulus. In the algorithmic fusion blocks, one image was randomly selected and always presented in the middle of the screen. In cognitive fusion blocks, the single-sensor trials required one image that was displayed either to the left or right of the center. In each cognitive fusion trial the stimuli were displayed with minimal visual angle to allow participants to keep their eyes fixated in the center of the screen without having to saccade for perceptual processing all of the information. Following the stimulus, a blank screen was presented for response. No trial-by-trial feedback was given.

To analyze differences in operator performance when presented with cognitive or algorithmic fusion, we first applied a traditional analysis of mean correct RTs and accuracy, followed by SFT analysis. For the SFT analysis, we estimated the capacity coefficient for each individual in each condition. We only analyzed the SIC and MIC of individuals for whom their data did not indicate a violation of selective influence. In the results section, we note whether a participant passed or failed selective influence. In order to pass selective influence we used paired Kolmogorov-Smirnov tests of RT survivor distributions to test that, for all t: $S_{HH}(t) < S_{HL}(t)$ and $S_{HH} \not> S_{HL}$, $S_{HH}(t) < S_{LH}(t)$ and $S_{HH} \not> S_{LH}$, $S_{LL}(t) > S_{HL}(t)$ and $S_{LL} \not< S_{HL}$, $S_{LL}(t) < S_{LH}(t)$ and $S_{HH} \not< S_{LH}$.

In Experiment 1 we investigated the processes underlying cognitive and algo-

DISTRIBUTION A: Distribution approved for public release.

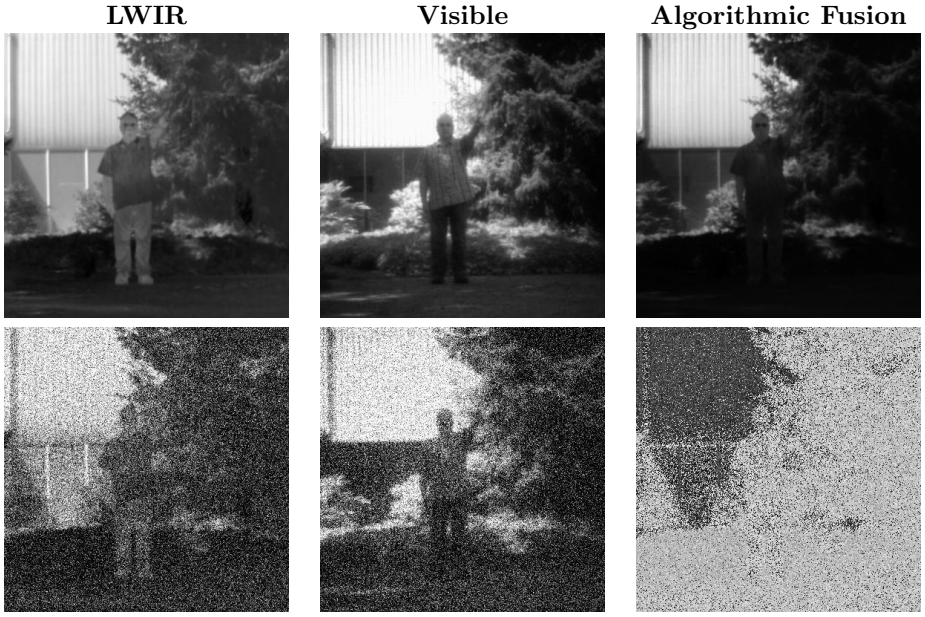

Figure 29: Examples of a LWIR, visible, and algorithmically fused image using the LPT algorithm both with (bottom images) and without (top images) white noise used for the pointing discrimination stimuli.

rithmic presentation of two related stimuli, those used for "pointing discrimination" and those used for "facing discrimination." Examples of each sensor image and the combined algorithmic image are shown in Figure 29 for the pointing discrimination and Figure 30 for the facing discrimination. We predicted that the facing discrimination stimuli would be more difficult than the pointing discrimination for two reasons: 1) the actor is always located in the center of the image for the pointing condition but in the facing condition the location of the actor varies across trials, and 2) the signal in the pointing discrimination stimuli (i.e., entire arm pointing left/right) is more salient than the signal in the facing discrimination stimuli (i.e., contours of the front versus back of the body). This prediction was supported by our findings.

Ten individuals (6 male, 4 female) participated in this study. Their ages ranged from 20 to 37 years (M = 25 years).

A total of $2 \times 2 \times 2 \times 10 = 80$ images were used in Experiment 1. There were two types of stimuli (pointing and facing), two sensor images (visible and LWIR) for each scene and an image could either indicate a person pointing (facing) to the "left" or "right." For each direction, there were 10 possible scenes (5 each of two people). See Figure 29 for example stimuli. Fusing the visible and LWIR pairs created an additional 40 images.

To reduce image salience, we added zero mean Gaussian luminance noise to the base image before displaying or fusing. Noise samples were independent

DISTRIBUTION A: Distribution approved for public release.

Figure 30: Examples of LWIR, visible, and algorithmically fused images used for the facing discrimination stimuli. In Experiment 1, white noise was added similar to Figure 29.

within and across images.

Each participant completed 5 days of 1-hour sessions for each stimulus type: pointing and facing (10 days total).

For the first set of stimuli (pointing), participants were asked to discriminate whether a person's arm was pointing left or right (see Figure 29). In the second set of stimuli (facing), participants indicated whether a person was facing toward to the left or the right side of the screen (see Figure 30). If the participant determined left, they pressed the left mouse button, if right, they pressed the right mouse button. The participants were told to perform the task as quickly and accurately as possible and were informed they must achieve at least 90% accuracy.

The first session of each stimulus type (Day 1: pointing, Day 6: facing) contained trials to compute the capacity coefficient for both cognitive and algorithmic fusion. Based on pilot data, simulations and time constraints, we collected 120 trials per image type needed for the capacity coefficient (LWIR-alone, visible-alone, LWIR and visible together). Hence, 360 trials were needed to estimate the capacity coefficient for cognitive fusion and 360 trials were needed to estimate the capacity coefficient for algorithmic fusion for a total of 720 trials.

Other sessions began with 120 trials dedicated to determining the noise level that would lead to 90% accuracy for each image type. This noise level was then used to for the low salience images in combination with the original images for the trials required to estimate the SIC. Based on pilot data, simulations and time constraints, we collected 270 trials per salience condition for a total of 1080 trials per session.

The sessions alternated between algorithmic and cognitive fusion (e.g., Day 2: cognitive fusion, Day 3: algorithmic fusion, Day 4: cognitive fusion, Day 5: algorithmic fusion).

Following the localization box, the stimulus was displayed for 250 msec. Whether the visible was on the right or the left was randomly varied in cognitive fusion trials. Following the stimulus, a blank screen was presented for 1750 msec

DISTRIBUTION A: Distribution approved for public release.

allowing the participant 2 seconds to respond starting from stimulus onset.

In summary, responses were faster and more accurate with visible imagery LWIR imagery for the pointing discrimination stimuli but the reverse is shown with a similar facing discrimination stimuli. Participants were limited capacity with both fusion types, more-so for algorithmic than cognitive fusion.

Because the number of sensors could not be fully crossed with fusion type (fused imagery, whether cognitive or algorithmic, included more than one sensor by definition) or with sensor-type (when two sensor types were present, then both IR and visible were necessarily displayed), we computed three separate repeated-measures ANOVA to examine, respectively, single to multi-sensor comparisons, within multi-sensor comparisons, and within single-sensor comparisons for the mean RT and accuracy.

Table 5 gives the results of a 2×2 repeated-measures ANOVA to assess effects of the number of sensors presented (single, multiple) and the stimuli (pointing, facing) for both correct RTs and accuracy. For both correct RTs and accuracy, there was a significant interaction between number of sensors and stimuli with main effects of the number of sensors presented and stimuli type (Table 5). Figure 31 indicates slower, less accurate performance with the facing discrimination stimuli. Across facing and pointing stimuli, performance with multi-sensor imagery suffers more than performance with single sensor imagery.

Table 6 gives the results of an additional 2×2 repeated-measures ANOVA to assess effects of the multi-sensor fusion method (algorithmic, cognitive) and the stimuli (pointing, facing) for correct RTs and accuracy. For correct RTs, we found a significant interaction between fusion method and stimuli type with a significant main effect of stimuli. However, we did not find a significant main effect of fusion method (likely due to the cross-over interaction). Analysis of accuracy (Table 6), indicated a significant interaction of fusion type and stimuli with significant main effects of both fusion type and stimuli. Figure 31 indicates algorithmic fusion is faster and slightly less accurate in the pointing discrimination, but is slower and less accurate in the facing discrimination.

Lastly, Table 7 gives the results of a $2 \times 2 \times 2$ repeated-measures ANOVA to assess the effects of single image presentation type (left/right of center, center), sensor (visible, LWIR), and stimuli (pointing, facing) to predict correct RTs and accuracy. For both correct RTs and accuracy, the three-way interaction of presentation type, sensor, and stimuli and two-way interaction between presentation type and sensor were not significant. For both correct RTs and accuracy there was a significant interaction of presentation type and stimuli and a significant interaction of sensor and stimuli with main effects of presentation type and sensor. There was a significant main effect of stimuli for correct RTs but not for accuracy.

Recall that for algorithmic fusion blocks, the single-sensor image was always presented in the middle of the screen. In cognitive fusion blocks, the single-sensor trials required one image that was displayed either to the left or right of the center. Figure 31 indicates both LWIR and visible single-sensor trials were faster and more accurate when visual attention was anticipating stimuli on a smaller visual area (algorithm-fused block of trials) than a larger visual area

DISTRIBUTION A: Distribution approved for public release.

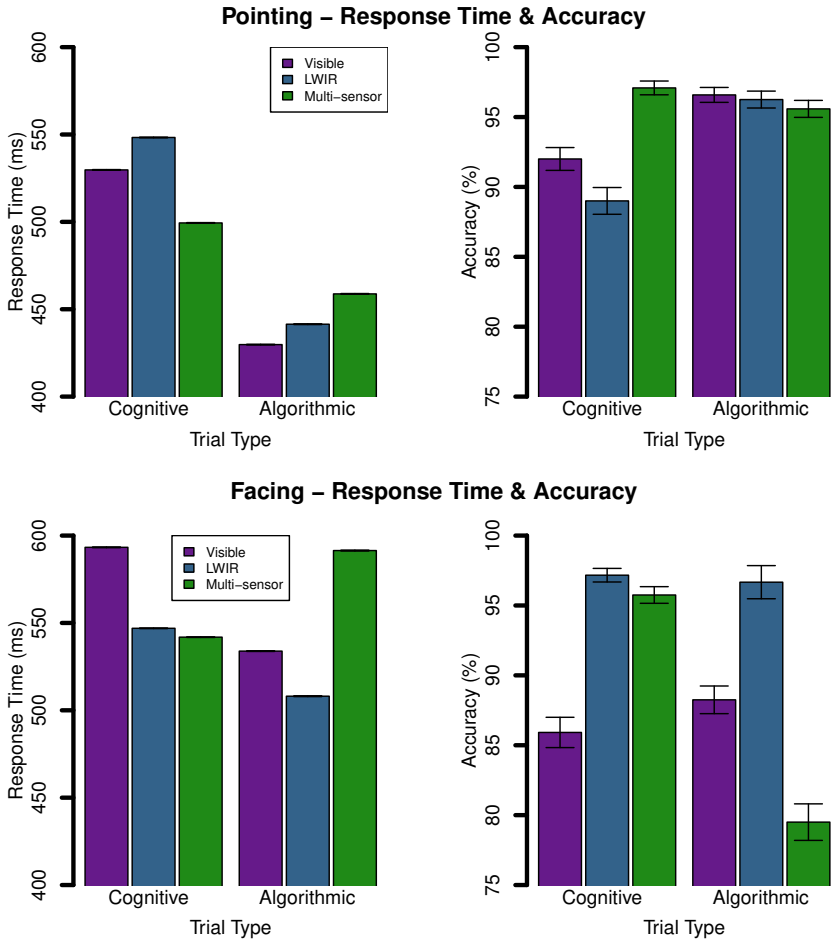

Figure 31: Mean correct RTs (left) and accuracy (right) for each sensor type for each fusion method in the pointing stimuli (top) and the facing stimuli (bottom): Cognitive fusion (visible and LWIR alone randomly presented on left/right of center) and algorithmic fusion (visible and LWIR alone presented in the center of the screen). Error bars represent the standard error of the mean (Jarmasz & Hollands, 2009).

DISTRIBUTION A: Distribution approved for public release.

(cognitive-fused block of trials) even though the same single-sensor image was presented in both conditions.

Further individual level analyses of the capacity coefficient and SIC allows us to examine how cognitive processing changes across the manipulated fusion type, sensor, and stimuli conditions by participant. Separate analyses of SFT were conducted for algorithmic and cognitive fusion across both pointing and facing stimuli for those who satisfy selective influence. We first report these results for the pointing stimuli, then the facing stimuli.

In the pointing stimuli, the capacity coefficient function was below 1 (i.e., limited capacity) for some time for both cognitive and algorithmic fusion for all participants. Individual capacity z-scores in the pointing stimuli ranged from -9.5 to -6.4 for algorithmic fusion and from -4.2 to 0.08 for cognitive fusion (Table 8). The performance hypotheses were supported at the group level: we found limited workload capacity across both fusion types (algorithmic fusion $t(9) = -28.36, p < .05, d = 12.68$ cognitive fusion $t(8) = -3.59, p < .05, d = 1.69$). Algorithmic fusion was significantly more limited than cognitive fusion ($t(8) = 8.54, p < .05, d = 3.99$).

For SIC analyses of cognitive fusion, selective influence could not be rejected for 6 participants based on a series of Kolmogorov-Smirnov (KS) tests. The Houpt-Townsend SIC statistic (Houpt & Townsend, 2010b) indicated 3 participants had a significant positive SIC, 1 participant had a significantly negative SIC and two participants had neither significant positive nor significant negative deviation, but did have a significantly positive MIC. Recall that a significance cutoff of $\alpha = .33$ was used for the SIC and MIC test. The remaining 4 participants failed tests of selective influence precluding the interpretation of their SICs. Table 10 lists each participant's Houpt-Townsend SIC statistic for both positive and negative deviations from zero, the MIC statistic, and the corresponding processing model.

Using the hierarchical Bayesian model we found minimal evidence for a zero MIC at the group level ($\hat{p}^0_{\text{posterior}} = .52$). The remaining models were unlikely ($\hat{p}^-_{\text{posterior}} = 0.19$; $\hat{p}^+_{\text{posterior}} = 0.29$). At the individual-level, the ratio of posterior odds (i.e., most likely model divided by the second-most likely model) did not show strong evidence of a particular processing architecture and stopping rule for any individual. The ratio of posterior odds ranged from 1.49 to 2.42. Note that using the Kass and Raftery (1995) scale, a ratio less than 3.2 is considered insufficient evidence from which to draw strong conclusions.

For algorithmically fused images, no participant's data satisfied the assumptions of selective influence thereby precluding the use of the SIC for model classification.

With the facing stimuli, Participant 1 did not obtain at least 80% accuracy in all conditions for further analysis of workload capacity with multi-sensor information. For other participants, $C(t) < 1$, for some time for both cognitive and algorithmic fusion. Capacity z-scores ranged from -10.7 to -8.5 for algorithmic fusion and from -4.9 to -2.2 for cognitive fusion (Table 9). We hypothesized that individuals' efficiency of both algorithmic and cognitive fusion was at least

DISTRIBUTION A: Distribution approved for public release.

as high as respective UCIP predictions (i.e., unlimited capacity) for the facing discrimination stimuli. The performance hypotheses were not supported at the group level; we found limited workload capacity (C(t) < 1) across both fusion techniques (algorithmic fusion $t(8) = -45.80, p < .05, d = 21.59$, cognitive fusion $t(9) = -14.32, p < .05, d = 6.40$) with algorithmic fusion significantly more limited than cognitive fusion, ($t(8) = 14.30, p < .05, d = 7.24$).

We divided individuals' data into 2 separate days to compute the SIC because no one participant passed the tests of selective influence when combining across days. For cognitive fusion SIC analyses, selective influence was not rejected for 4 participants for 1 of the 2 days of data collection. All 4 participants' SIC function had no significant deviations from zero. Table 11 lists each participant's Houpt-Townsend SIC statistic for both positive and negative deviations from zero, the MIC statistic, and the processing model that would predict that pattern of significance.

Using the hierarchical Bayesian model we found minimal evidence for a zero MIC at the group level ($\hat{p}^0_{\text{posterior}} = .54$). The remaining models were unlikely ($\hat{p}^+_{\text{posterior}} = 0.28$; $\hat{p}^-_{\text{posterior}} = 0.18$). All participants' most likely model was MIC $= 0$ and the second-most likely model MIC > 0. For these participants, the ratio of posterior odds ranged from 1.62 to 2.63 indicating very weak evidence for each individual. Thus for both individual- and group-level conclusions we found weak evidence for a serial processing model. These results are consistent with SIC findings of no significant deviations from zero.

As with the pointing, algorithmically fused images, no participant's data satisfied the assumptions of selective influence thereby precluding the use of the SIC for model classification.

We used a repeated-measures ANOVA to examine the effects of stimulus (pointing, facing) and fusion type (cognitive, algorithmic) on capacity z-scores. The interaction was non-significant, $F(1, 8) = 0.03, p = 0.87, \eta^2_G = 0.00$, and the main effect was significant for both stimulus type, $F(1, 8) = 20.53, p < .05, \eta^2_G = 0.37$, and fusion type, $F(1, 8) = 137.94, p < .05, \eta^2_G = 0.87$. Capacity z-scores with the pointing stimuli were higher than z-scores with the facing stimuli for both fusion types, with algorithmic fusion more limited than cognitive fusion.

Both cognitive and algorithmic fusion hindered processing of the individual source images relative to independent parallel processing. Because information was redundant across the two images, participants should be faster with two images than with a single image, even with independent parallel processing of each image (cf. Raab, 1962). Subjects were slightly faster with the side-by-side images than the single sources images, however the capacity results indicate that the speed-up was not as much as would be observed from independent parallel processing. Performance was even worse with the algorithmically fused images: RTs were slower with algorithmically fused images than with either of the single sensor images. Hence, capacity coefficient values were quite low for algorithmic fusion, much lower than cognitive fusion.

Low capacity coefficient values can result from a number of different violations of the baseline UCIP model predictions. All other factors being equal,

DISTRIBUTION A: Distribution approved for public release.

serial processing systems are more limited capacity than parallel, while coactive processing systems have higher capacity than standard parallel (Townsend & Nozawa, 1995b; Townsend & Wenger, 2004b).[9] Unfortunately, our results from the SIC analysis did not lead to clear results regarding processing architecture. All participants' data indicated violations of selective influence for the algorithmically fused images. Most participants indicated a violation of selective influence with cognitive fusion. Of those participants that did not violate the distribution ordering implied by selective influence, null-hypothesis testing indicated a variety or processing strategies: parallel-OR process and parallel-AND with the pointing stimuli and serial-OR with the facing stimuli. The Bayesian analysis of the MIC indicated that there very slight evidence in favor of a zero MIC a the group level (MIC = 0) and similarly minimal evidence for any MIC category (positive, negative or zero) at the individual level for both stimuli types.

Among those participants that may be using a parallel-OR processing strategy, capacity coefficients were still quite limited indicating that there may be other deficits relative to the UCIP model. Given the short presentation time and the fact that at least one of the images was extrafoveal, a violation of the "unlimited capacity" assumption is a likely cause. With a single image, participants can fixate on the most informative region of that image to get the most out of the image. When there are two images, at most one can be fixated so information uptake is almost certainly not the same with two images relative to one. Limitations of visual short-term memory may degrade the ability to integrate information from multiple sensors or potentially facilitate the strategy to only process a single, informative sensor image (Irwin, 1991; Rayner et al., 1980).

With algorithmic fusion, only a single image is presented, so participants can fixate the most informative region. Hence, the limitations on visual attention that may explain low capacity values for cognitive fusion are not sufficient for algorithmic fusion. Although we were not able to draw direct inferences from the SIC, we can make some inferences about the processing. Independent serial or parallel processing are unlikely candidates, as they should have led to effective selective influence and hence ordered distributions Dzhafarov (2003); Houpt, Blaha, McIntire, Havig, and Townsend (2014); Houpt and Townsend (2010b). A priori, it is difficult to imagine how (or why) the visual system would separate the information from each source before processing. Indeed, previous research using sophisticated accuracy based methodologies found that individual sensor information was perceptually nonseparable in an algorithmically combined image (McCarley & Krebs, 2006). Because the combined algorithmic image is processed as a single unit of information that integrates information from both sensors, the visual processing decision is similar to a coactive process. However, unlike most coactive processes, the capacity values are much lower than independent parallel, not higher. This suggests that there is useful information lost in

[9] In fact, some authors define coactive processing by violations of the race model inequality, an upper bound on parallel processing with context invariance (cf. Miller, 1982).

DISTRIBUTION A: Distribution approved for public release.

the fusion process, perhaps more akin to an inhibitory parallel process (cf. Eidels et al., 2011b). The potential information loss is evident in Figures 29 and 30, in which the person looks more clearly differentiated from the background in the single sensor images than in the algorithmically fused image.

Based on McCarley and Krebs (2000) and Krebs and Sinai (2002), we had assumed that a more difficult stimulus set (i.e., degraded quality of image, type of psychophysical task) would lead to higher capacity coefficient values for the algorithmically fused imagery. The more difficult stimuli in our experiment was based on the facing stimuli were not always directly centered (as the pointing stimuli were centered) and there were fewer physical cues to aid in decision-making. Capacity was higher at the group level with the pointing stimuli than with the facing stimuli when using algorithmic fusion (as well as cognitive fusion), although it was not enough of an increase to reach the capacity values from cognitive fusion, let alone the predicted UCIP baseline.

There was some evidence of a differential speed-accuracy trade-off between the algorithmically fused imagery and the cognitively fused images. Algorithmic fusion led to faster and slightly less accurate performance than cognitive fusion in the pointing stimuli. However, algorithmic led to both slower and much less accurate performance than cognitive fusion in the facing stimuli. This may suggest that different fusion approaches may be more appropriate for situation in which accuracy or speed are more important, at least for more simple discriminations, but more exploration is necessary.

Differences in speed-accuracy focus can be problematic for capacity coefficients. Assessment functions (Donkin, Little, & Houpt, 2014; Townsend & Altieri, 2012) are a variation on the capacity coefficient that can ameliorate this problem, however there are not inferential statistics available for the assessment function so we only reported capacity coefficients. We did calculate assessment functions and in all cases, the visual patterns matched our conclusions drawn from the capacity coefficients. These data indicate no significant speed-accuracy impact on processing efficiencies for either algorithmic or cognitive fusion.

In Experiment 1, we obtained clear results indicating limited capacity for extracting information from multi-sensor imagery, with both cognitive and algorithmic fusion. The results regarding architecture were less clear and our goal in Experiment 2 is to obtain more robust results from the SIC and MIC analyses. There are a number of potential reasons for the variability across subjects in the SIC results and the relatively weak evidence indicated by the MIC test. First, many participants' data was not usable due to the lack of survivor function ordering that is necessary for SIC analyses. This meant that there were very few SIC/MIC combinations available from which to draw conclusions. Hence, we doubled the number of participants for Experiment 2. Second, participants in Experiment 1 may not have settled on a particular strategy and hence their data may represent a mixture of parallel and serial processing. To address this issue, participants in Experiment 2 had 8 days of experience with the single and fused imagery before we collected data for the SIC/MIC. Furthermore, we limited the stimuli to the facing stimuli stimuli from Experiment 1.

For the 8 days of training we added noise to every LWIR and visible image

DISTRIBUTION A: Distribution approved for public release.

to slow-down the processing of the image information and allow for improvements in performance over the course of training as more efficient strategies may develop over training. We did so because in Experiment 1 participants demonstrated similar correct RTs in single sensor conditions (LWIR-only, visible-only) and multi-sensor conditions across both algorithmic and cognitive fusion presentations without any kind of training, strategy instructions, and only brief stimulus presentation times. Therefore, we wanted to slow processing down to leave room for further possible improvements of supplying multiple sensors and several days of training.

In place of the Gaussian white noise used in Experiment 1, we added pink noise instead of white noise (which we had used in Experiment 1) for more naturalistic degradation of image quality (Glasgow et al., 2003; Reis, Marasco, Havig, & Heft, 2004; Reis et al., 2004). Example stimuli are shown in Figure 32.

Finally in Experiment 2, we only measured the SIC/MIC for cognitive fusion. Although we did measure capacity coefficients for both cognitive and algorithmic fusion, we did not further examine algorithmic fusion method because results from Experiment 1 indicated that selectively influencing each source image would be unlikely if not impossible.

We expected participants to exhibit higher accuracy and lower correct RTs with training. The capacity coefficient represents an improvement in RTs relative to the improvement with single source images. If training affects not only the perception of each source, but also the efficiency with which participants use the combined information, then we would also expect capacity to increase over training. Alternatively, if there is no additional improvement for the process of combining the information, then the capacity would be stable across training.

Additionally, we hypothesized that participants would use a consistent strategy after training, hence correct RTs would indicate a clear SIC signature (see Figure 1) and strong evidence from their MIC.

Twenty individuals (12 male, 8 female) participated in this study. Their ages ranged from 21 to 34 years (M = 24 years).

Stimuli were selected from Experiment 1 from the facing discrimination stimuli. We chose to use only the actor who participants from Experiment 1 indicated was the most clear across the images. To increase the size of the base image set and control for extraneous variation in the images, we edited the images to manipulate the direction the actor was facing and the spatial location of the actor in the image. The editing process involved placing the LWIR and visible image of the actor in 10 locations across the image scene. The background scene was averaged across all images to avoid any distortion or aberrations that could influence participant performance. In total, there were 160 stimuli: 2 sensors (LWIR, visible) $\times$ 2 directions (left, right) $\times$ 2 backgrounds (raw, inverted) $\times$ 2 poses (standing, snapshot while walking) $\times$ 10 locations (various, ecologically valid, placements across the image). One LWIR-visible pair (same direction, background, pose, and location) was randomly selected for each trial. The LWIR-visible pairs were algorithmically fused to create 80 additional stimuli.

The amount of pink noise was consistent during training within and across

DISTRIBUTION A: Distribution approved for public release.

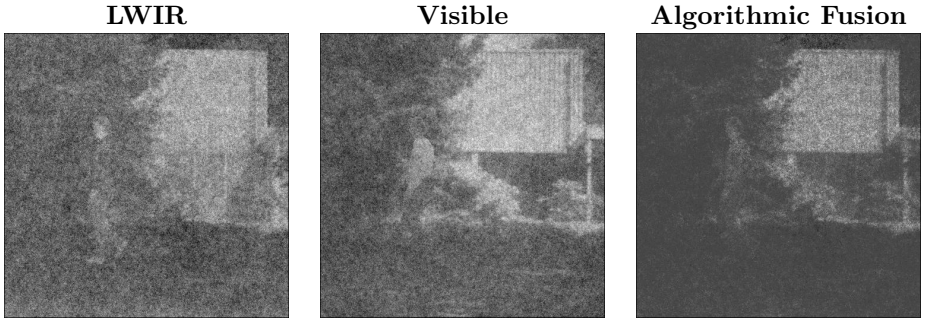

Figure 32: Examples of LWIR, visible, and algorithmically fused images used for the facing discrimination stimuli. In Experiment 2, pink noise was added to every image during the 8 training sessions.

participants. We targeted 82% accuracy for each source using the Quest psychometric estimation method with pilot subjects. We chose 82% because it leads to 96% overall accuracy in a UCIP system $(1 - (1 - 0.86)^2 = 0.96)$.

Experiment 2 instructions were the same as those used with the facing stimuli in Experiment 1. Participants indicated whether a person was facing toward to the left or the right side of the screen (see Figure 30) using the corresponding mouse button. Participants were told to perform the task as quickly and accurately as possible. At the end of each session, participants were informed of their accuracy in each fusion condition. This feedback was provided to keep participants motivated to improve in performance over the course of training sessions.

Each participant completed 10 days of 1-hour sessions. The first 8 sessions contained trials to compute the capacity coefficient for both cognitive and algorithmic fusion. As with Experiment 1, there were 120 trials per distribution (LWIR-alone, visible-alone, LWIR and visible together) for a total of 720 trials for capacity analysis.

The remaining two sessions required first the estimates of each sensors psychophysical thresholds at 82% accuracy by manipulating the amount of pink noise added to the image (120 trials each sensor, each day) followed by trials required to estimate the SIC (2160 trials total). The SIC trials consisted of factorial combinations of high (no noise) and low (individualized amount of pink noise) of both the LWIR and visible images. LWIR was always presented on the left, visible on the right. For trials with only one sensor present (e.g., LWIR with high salience, visible is absent) the localization box would appear in place of the image (example shown in Figure 33).

In cognitive fusion blocks, we fixed the location of where the LWIR and visible images are presented across all trials (LWIR=left of center, visible=right of center) instead of randomly displaying each on the left/right for every trial (as in Experiment 1). This gave operators the opportunity to anticipate where each type of information was going to be presented.

DISTRIBUTION A: Distribution approved for public release.

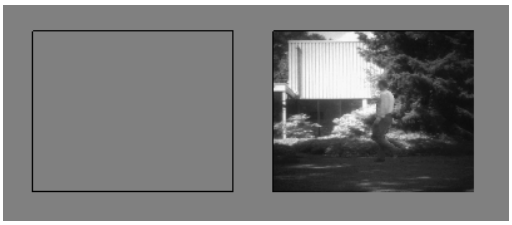

Figure 33: Example of a cognitive fusion presentation of LWIR (absent) and visible (high). The participants were asked to discriminate whether the person was facing to their left or right. The two images were centered and presented within 6.39°of visual angle on a mid-gray background.

Stimulus presentation duration was extended to 2 seconds across all conditions (algorithmic and cognitive, single- and multi-sensor) to allow the operator to sample all of the information from each image and allow strategies of processing the information to potentially improve with time. The LWIR was always displayed on the left and the visible image was always displayed on the right. Following the stimulus, a blank screen was presented for 500 msec allowing the participant total of 2.5 seconds to respond starting from stimulus onset.

RTs and accuracy with fused imagery was worse than single-sensor images. Performance on both single and multi-sensor imagery improved with training, however the capacity coefficient consistently indicate inefficient performance with bouth algorithmic and cognitive fusion, lower capacity results for algorithmic fusion than cognitive fusion, and no efficiency improvements with training. Nonetheless, we found strong evidence for parallel and coactive processing strategies with cognitive fusion, both of which are normally associate with efficient processing.

Table 12 gives the results of a 2×8 repeated-measures ANOVA to assess effects of the number of training sessions completed and the type of fusion (algorithmic, cognitive) for both correct RTs and accuracy for trials with multiple sensors. There was an interaction between training and fusion technique in accuracy, but not RT. For both correct RTs and accuracy, we found a main effect of the number of training sessions completed. There was not a main effect of fusion technique (algorithmic, cognitive) for correct RTs, but there was for accuracy.

Although performance clearly improves over training, it is not clear if the efficiency with which individuals use the fused imagery improves from the mean RT and accuracy data. For this information, we need the capacity results which are presented in the next subsection.

Table 13 gives the results of a 2×8 repeated-measures ANOVA to analyze the effects of training on the efficiency of processing multi-sensor information to predict capacity z-score values. Participant 12 and 17 were excluded from efficiency comparisons across training sessions because of low accuracy in the early training sessions. Figure 35 illustrates that individual capacity

DISTRIBUTION A: Distribution approved for public release.

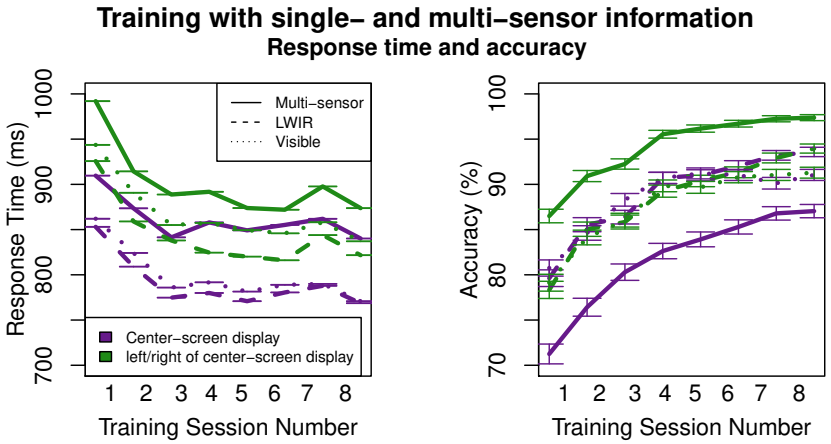

Figure 34: Group-level means of correct RTs and accuracy across days of training. Line type indicates the type of imagery used: fused (solid), LWIR (dashes), or Visible (dots). Line color indicates the screen layout of the images: single center-screen images (purple), or left/right/both images (green). Hence, the algorithmic fusion results (multi-sensor, center-screen) are indicated by solid purple lines and the cognitive fusion results (multisensor, left/right of center) are indicated by solid green lines. Error bars represent the standard error of the mean (Jarmasz & Hollands, 2009).

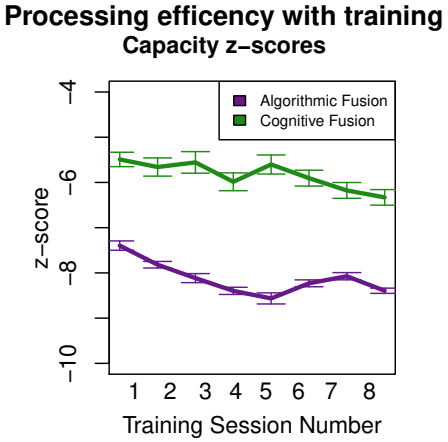

Figure 35: Group-level means of capacity z-scores representing the processing efficiency of multi-sensor information computed for each day of training with both algorithmic and cognitive fusion. Error bars represent the standard error of the mean (Jarmasz & Hollands, 2009).

DISTRIBUTION A: Distribution approved for public release.

z-scores with cognitive fusion were less limited than z-scores with algorithmic fusion ($t(143) = -12.19, p < .05, d = 1.45$). Capacity z-scores become significantly more limited from the first (Day 1) to last (Day 8) day of training for both algorithmic fusion ($t(17) = 3.03, p < .05, d = 1.02$), and cognitive fusion, ($t(17) = 2.99, p < .05, d = 0.49$).

Table 14 indicates the participants whose data passed the selective influence test, the participants' Houpt-Townsend SIC statistic for both positive and negative deviations from zero, the MIC statistic, and the processing model that predicts their pattern of results. Distributional orderings did not indicate violations of selective influence for 11 participants. Ten of those participants had a significantly ($p < .33$) positive SIC. Four of those participants had a significantly positive MIC and significant negative SIC. One participant had a significantly positive and negative SIC with a non-significant MIC. One participant had a significantly negative SIC. Participant 9 had SIC/MIC results that are not predicted by any of the independent serial/parallel/coactive AND/OR models.

With the hierarchical Bayesian MIC analysis, we found good evidence for a positive MIC at the group level ($\hat{p}^{+}_{\text{posterior}} = 0.73$). The remaining models were equally unlikely ($\hat{p}^{-}_{\text{posterior}} = 0.15$; $\hat{p}^{0}_{\text{posterior}} = 0.12$). At the individual-level, the posterior probabilities supported the conclusions drawn from the Houpt-Townsend statistic of positive and negative deviations of the SIC (Table 14). Nine participants' most likely model had MIC > 0 with MIC $= 0$ as the second-most likely. Among those participants, the ratio of posterior odds ranged from 4.5 to 70.0 indicating strong to decisive evidence for each individual. Participant 9 had a most likely positive MIC with MIC < 0 second-most likely. Participant 20 had a most likely negative MIC with MIC > 0 second-most likely. Both Participant 9 and 20 had minimal evidence in favor the most likely model, with a ratio of posterior odds of 1.9 and 2.0 over the next best model respectively.

In Experiment 2 our aim was to produce consistency within an individual and across people in the processes involved with multi-sensor information. We found nearly identical capacity results with those of Experiment 1 despite the several experimental changes: 1) increased experience with multi-sensor imagery, 2) realistic degradation of image quality with pink noise, 3) longer stimulus presentation time, and 4) fixing LWIR to left-hand side of the screen and visible to the right-hand side. Even with many experimental changes we consistently found limited workload capacity with both algorithmic and cognitive fusion. Similarly, the discrepancy between single- and multi-sensor performance with algorithmic fusion was much larger than cognitive fusion. Likewise, we found lower capacity results for algorithmic fusion than cognitive fusion.

When participants had undergone training, there clear results indicating processing architecture from SIC analyses. We found group-level evidence of parallel-OR or coactive processing (the MIC cannot distinguish between these processing strategies). The ability to process both images in parallel leaves to opportunity for facilitation in performance from the redundancy speed-ups across the two images (Kahneman, 1973; Pollatsek et al., 1984).

Over the course of training performance improved for all single and multi-

DISTRIBUTION A: Distribution approved for public release.

sensor conditions. These raw RT results cannot discriminate whether the multi-sensor performance improvement was due to better use of single-sensor images or improvements in the *integration* of the sensor images. By applying the capacity coefficient, it was clear that integration of multi-sensor imagery did not improve with training, and in fact may have degraded.

Despite limited capacity results, we still find evidence for efficient processing strategies. SIC and MIC results from the cognitive fusion conditions indicate clear evidence against serial processes, in favor parallel-OR or even coactive processing. Although we could not draw conclusions from the algorithm-fused imagery, we assumed serial processing of each source was highly improbable, and the process is more likely a type of coactivation. Thus, the limited capacity results are not due to inefficient serial processing of information. For cognitively-fused imagery, the available processing capacity could be divided between the two sources of information and in turn slow down the processing of the individual sensors or the information provided from each sensor inhibits processing of the alternative. For algorithm-fused imagery, limited capacity results may result from inhibition that degrades sensor integration in the overall composite image.

Across two experiments we found strong evidence of limited capacity for both algorithmic and cognitive fusion. Although in some cases, RTs were faster with fused imagery, they were not as fast as our model predicted given that the redundant information across the two sources. Despite the mixed effects we found with raw RTs, the capacity coefficient indicated algorithmic fusion led to more limited capacity performance than cognitive fusion, despite only requiring participants to attend to one image. These capacity results were consistent across a variety of manipulations: stimuli (facing, pointing), difficulty (no noise, pink noise), viewing duration, and variability in single sensor image placement on the screen (random, predictable).

Image fusion may have the best results when each sensor alone does not supply redundant information; rather, only the configural combination of the information allows for correct decision-making (Klein et al., 2006; Neriani et al., 2008). For instance, Toet et al. (1997) found performance improvements with algorithmically fused LWIR and visible images, contradictory to our findings. The task used in Toet et al. (1997) was tailored to specifically utilize both visible and LWIR information. The participants were asked to determine the position of a person relative to an environmental object (i.e., fence, walkway, or tree). Therefore, to correctly identify the spatial location the participant must take advantage of unique information from each sensor. Follow up studies should consider performance comparisons across multi-sensor information presented with algorithmic and cognitive fusion when the individual sensors each supply unique, useful information to the observer.

In many cases, it may be difficult to determine a priori the extent to which task-relevant information is redundant across sensors. There is some promise in the recent work by Bittner (2015) which uses response classification (e.g., Ahumada, 2002; Ahumada & Lovell, 1971) to assess the unique information used to make a decision from each sensor image. Response classification uses noise masking to identify the useful information in each single-sensor and multi-

DISTRIBUTION A: Distribution approved for public release.

sensor image for an observer to make a decision. Clusters of pixels can determine what unique features of each image carry task relative details.

Based on the existing research with algorithmic image fusion, we expected fusion would provide, at a minimum, equally efficient processing as an unlimited capacity, independent and parallel processing model. However, our results indicate just the opposite in it has been an assumption for multi-spectral fusion to enhance both speed and accuracy performance compared to individual sensor images. This discrepancy is partially due to alternative methods of analysis. For some conditions, the traditional analyses of RTs would indicate a benefit in performance with cognitive fusion compared to either single-sensor alone (Figure 31). While it seems as if performance is enhanced with the side-by-side presentation, these RT speed-ups are not faster than what can be attributed to what is expected when completing a task that only demands one source and the fastest of the two can be sampled on each given trial (i.e., statistical facilitation; Raab, 1962).

Some previous research based on traditional analyses has suggested that algorithmic fusion, at best, performed just as well as individual sensor performance and potentially hinders performance or situational awareness (Krebs, Scribner, Miller, Ogawa, & Schuler, 1998; Krebs & Sinai, 2002; Steele & Perconti, 1997). In those studies and our current work, it is possible that the quality of information in the algorithmically fused image was degraded compared to the individual sensor images. Even if the fused image were of equal quality to one or the other of the original images, it would not be sufficient to achieve unlimited capacity performance because there would be no opportunity for redundancy gain. The algorithmically fused image would need to have *better* information quality than either single-source image.

The potential reduction in image quality may be due to the fact that no consideration of the task or stimuli was used in choosing the particular algorithm. If task-specific image enhancement techniques are not utilized, task-relevant information may be filtered out in the fusion (Dixon et al., 2006; Toet & Hogervorst, 2012). Ideally, the choice of algorithm should attempt to adjust to particular task demands and environmental constraints to obtain optimal scene information, (e.g,. Yong, Weiqi, & Rui, 2010), however when systems are designed for general use, the task many not be known in advance.

For cognitive fusion, we found RT speed-ups for some conditions when comparing an individual sensor image to the presentation of both images side-by-side. However, those speed-ups were not significantly faster than our predicted model baseline. A limited capacity may result from any violation of the baseline assumptions: unlimited capacity, independence, or parallel processing. Using careful experimental control in Experiment 2 we saw strong evidence for parallel (even coactive) processing, leaving two potential explanations for limited capacity with cognitive fusion. Although the capacity coefficient cannot directly distinguish between violations of independence and workload we can speculate about the potential underlying mechanisms using previous research in conjunction to our findings: 1) There could be a limitation of workload capacity, or 2) there could be dependencies between processing of the two sources of informa-

DISTRIBUTION A: Distribution approved for public release.

tion (Eidels et al., 2011b). Although the first is possible, there would have to be an extreme workload capacity limitation to overcome the benefits of coactivation (cf. Townsend & Wenger, 2004b). In favor of the latter, McCarley and Krebs (2006) used general recognition theory (GRT; Ashby & Townsend, 1986) and found the perceptual dimensions of algorithmically combined imagery are nonseparable. In future research we are interested in investigating cognitive fusion with GRT as well.

We demonstrated that SFT aids in assessing various display alternatives by providing additional information about how an operator processes the information in each comparison of interest. We found strong evidence for limited capacity processing of both algorithmic and cognitive fusion of multi-sensor imagery. Despite requiring attention to only a single, composite image, algorithmic fusion resulted in more limited capacity than cognitive fusion across several experimental manipulations. Algorithmic fusion may only be beneficial when particular image preprocessing techniques can maximize the strengths of the algorithm given the stimulus environment.

While training participants with the task and imagery can reduce response times and increase accuracy for both single-source images and algorithmically or cognitively fused images, the efficiency with which participants combine the information does not improve. This lack of efficiency improvement was evident with both algorithmic and cognitive fusion. Despite the consistent inefficiency, individuals can simultaneously process multiple sensor images in parallel.

For unknown task environments, presenting all of the information to the operator gives them the opportunity to decide what is useful given the task. However, multi-sensor display may only be beneficial when each single-sensor provides unique, useful information to contribute to correct decision-making. System designers should not eliminate the potential for using display methods that provide all of the information while minimizing the operators invested attentional resources.

DISTRIBUTION A: Distribution approved for public release.

	Correct RT			Accuracy		
Condition	F	df	η^2_G	F	df	η^2_G
# of sensors × Stimuli	12.45**	1,9	0.01	60.53***	1,9	0.19
# of sensors	5.28*	1,9	0.00	9.54*	1,9	0.05
Stimuli	11.57**	1,9	0.28	17.19**	1,9	0.36

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; η^2_G: Generalized eta-squared

Table 5: Experiment 1 ANOVA results for the number of sensor images (1 or 2; visible and LWIR sensors presented alone or simultaneously) and the experimental task (pointing, facing) predicting correct RTs and accuracy.

	Correct RT			Accuracy		
Condition	F	df	η^2_G	F	df	η^2_G
Fusion technique × Stimuli	6.73*	1,9	0.09	56.84***	1,9	0.43
Fusion technique	0.07	1,9	0.00	76.17***	1,9	0.52
Stimuli	13.87**	1,9	0.27	37.62***	1,9	0.51

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; η^2_G: Generalized eta-squared

Table 6: Experiment 1 ANOVA results for the type of fusion technique used to combine the visible and LWIR images (cognitive or algorithmic fusion) and the experimental stimuli (pointing, facing) predicting correct RTs and accuracy.

5.3 Human Orientation Discrimination in Video

In our previous experiments, we explored how people combine several types of multispectral imagery using SFT. Using SFT allowed us to not only understand if performance differed, but also provided information about why performance is better or worse. We compared a display that provides each image beside one another and a display that provides a single image comprised of multiple spectral images. We found that overall processing efficiency of the information in each image declined as we moved from displaying a single-image to multiple images, regardless of how we combined the information together. Additionally, we found that providing an operator with all of the information, side-by-side, allowed them to perform just as well as when the images are algorithmically combined together, which traditionally is thought of the best method of displaying multiple images simultaneously. SFT gave us the appropriate framework to compare the processing of single- and multi-image information displays across the two types. We were able to narrow down what processing strategies individuals were showing using SFT results. We could eliminate the possibility of people sequentially processing each image as the cause of performance decrements using results of SFT. We found participants were able to processing both images simultaneously and sometimes fully-integrated.

The application and understanding of static imagery is important, but dynamic fusion is usually provided of real-world scenarios. A designer may have concerns about whether the results found of processing strategies for static stim-

DISTRIBUTION A: Distribution approved for public release.

	RT			Accuracy		
Condition	F	df	η^2_G	F	df	η^2_G
Display method × Sensor × Stimuli	5.11	1,9	0.00	2.77	1,9	0.01
Display method × Sensor	2.04	1,9	0.00	0.11	1,9	0.00
Display method × Stimuli	6.92*	1,9	0.04	7.29*	1,9	0.07
Sensor × Stimuli	32.58***	1,9	0.05	32.56***	1,9	0.28
Display method	92.53***	1,9	0.26	6.46*	1,9	0.08
Sensor	6.15*	1,9	0.00	40.46***	1,9	0.13
Stimuli	8.93*	1,9	0.22	2.11	1,9	.04

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; η^2_G: Generalized eta-squared

Table 7: Experiment 1 ANOVA results for the method used to present the single sensor image to the observer (center of the screen or randomly set to the left or right of center screen) and the type of sensor (visible, LWIR) and the experimental stimuli (pointing, facing) predicting correct RTs and accuracy.

	Algorithmic		Cognitive	
Subject	**Capacity**	**z-score**	**Capacity**	**z-score**
1	Limited	-8.174***	N/A	N/A
2	Limited	-6.367***	Unlimited	-0.088
3	Limited	-8.182***	Unlimited	-0.653
4	Limited	-7.694***	Limited	-4.056***
5	Limited	-7.780***	Unlimited	0.088
6	Limited	-9.155***	Limited	-3.322***
7	Limited	-7.436***	Limited	-4.219***
8	Limited	-7.547***	Limited	-4.066***
9	Limited	-7.660***	Limited	-2.362*
10	Limited	-9.500***	Unlimited	-0.826

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 8: Experiment 1: Individual level capacity, z-score, and statistical significance for algorithmic and cognitive fusion of multi-sensor images compared to UCIP model in the pointing discrimination stimuli.

DISTRIBUTION A: Distribution approved for public release.

	Algorithmic		Cognitive	
Subject	**Capacity**	**z-score**	**Capacity**	**z-score**
1	N/A	N/A	Limited	-3.992
2	Limited	-9.586***	Limited	-3.985***
3	Limited	-9.137***	Limited	-3.985***
4	Limited	-8.597***	Limited	-4.757***
5	Limited	-9.702***	Limited	-4.879***
6	Limited	-10.748***	Limited	-3.459***
7	Limited	-9.517***	Limited	-4.515***
8	Limited	-8.980***	Limited	-4.189***
9	Limited	-10.036***	Limited	-4.296***
10	Limited	-9.750***	Limited	-2.676**

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 9: Experiment 1: Individual level capacity, z-score, and statistical significance for algorithmic and cognitive fusion of multi-sensor images compared to UCIP model in the facing discrimination stimuli.

Subject	**Pass/Fail**	**D+**	**D-**	**MIC**	**Architecture**
4	Pass	0.018	$\mathbf{0.131}^{+}$	$\mathbf{-61.912}^{+}$	Parallel-AND
5	Pass	$\mathbf{0.180}^{+}$	0.065	**15.943***	Parallel-OR
6	Pass	$\mathbf{0.179}^{+}$	0.055	$\mathbf{9.09}^{+}$	Parallel-OR
8	Pass	$\mathbf{0.159}^{+}$	0.073	$\mathbf{25.321}^{+}$	Parallel-OR
9	Pass	0.096	0.086	$\mathbf{12.667}^{+}$	Ambiguous
10	Pass	0.101	0.011	$\mathbf{37.663}^{+}$	Ambiguous

Note: **H-T Statistic** = $^{+}p < 0.33$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.
MIC = $^{+}p < 0.33$, * $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

Table 10: Cognitive fusion results of the pointing stimuli in Experiment 1 including: Whether the participant (for a particular day) passed the test of selective influence, the Houpt-Townsend statistic (D+, D-), the mean interaction contrast (MIC), and the identified processing model. Bold D+ and D- statistics indicate a significant Houpt-Townsend statistic at $p < 0.33$.

DISTRIBUTION A: Distribution approved for public release.

Subject	Pass/Fail	D+	D-	MIC	Architecture
6.2	Pass	0.154	0.075	16.839	Serial-OR
7.2	Pass	0.136	0.123	40.692	Serial-OR
8.1	Pass	0.118	0.110	6.251	Serial-OR
9.2	Pass	**0.192**$^{+}$	0.069	4.310	Ambiguous

Note: **H-T Statistic** = $^{+}p < 0.033$, * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.
MIC = $^{+}p < 0.033$, * $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$.

Table 11: Cognitive fusion results of the facing stimuli in Experiment 1 including: Whether the participant (for a particular day) passed the test of selective influence, the Houpt-Townsend statistic (D+, D-), the mean interaction contrast (MIC), and the identified processing model. Bold D+ and D- statistics indicate a significant Houpt-Townsend statistic at $p < 0.33$.

	Correct RT			**Accuracy**		
Condition	F	df	η^2_G	F	df	η^2_G
# of training sessions × Fusion technique	2.05	7,133	0.01	2.37*	7,133	0.02
# of training sessions	5.03***	7,133	0.05	19.92***	7,133	0.32
Fusion technique	2.14	1,19	0.02	329.18***	1,19	0.49

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; η^2_G: Generalized eta-squared

Table 12: Experiment 2 ANOVA results for the number of training sessions (1-8) and the fusion technique (algorithmic, cognitive) predicting correct RTs and accuracy for multi-sensor trials.

	z-score		
Condition	F	df	η^2_G
# of training sessions × Fusion technique	0.03	1,16	0.00
# of training sessions	10.29 **	1,16	0.05
Fusion technique	21.12***	1,16	0.53

Note: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$; η^2_G: Generalized eta-squared

Table 13: Experiment 2 ANOVA results for the number of training sessions (1-8) and the fusion technique (algorithmic, cognitive) predicting group-level mean capacity z-scores.

DISTRIBUTION A: Distribution approved for public release.

Subject	Pass/Fail	D+	D-	MIC	Architecture
3	Pass	**0.349*** **	**0.241** **	**60.101*** **	Coactive
9	Pass	**0.160**$^{+}$	**0.182***	**−4.752**$^{+}$	Ambiguous
10	Pass	**0.190**$^{+}$	0.071	**48.077**$^{+}$	Parallel-OR
11	Pass	**0.257** **	**0.125**$^{+}$	**103.470***	Coactive
13	Pass	**0.429*** **	0.071	**152.638*** **	Parallel-OR
14	Pass	**0.109**$^{+}$	0.151	16.710	Ambiguous
15	Pass	**0.263** **	**0.225***	**51.046***	Coactive
16	Pass	**0.230***	**0.154**$^{+}$	**51.050**$^{+}$	Coactive
17	Pass	**0.198***	0.048	**62.970*** **	Parallel-OR
19	Pass	**0.142**$^{+}$	**0.258** **	32.772	Serial-AND
20	Pass	0.041	**0.165**$^{+}$	-42.617	Ambiguous

Note: **H-T Statistic (D+, D-)** $= {}^{+}p < 0.33$, * $= p < 0.05$, ** $= p < 0.01$, *** $= p < 0.001$.
MIC $= {}^{+}p < 0.33$, * $p < 0.05$, ** $= p < 0.01$, *** $= p < 0.001$.

Table 14: Cognitive fusion results of Experiment 2 including: Whether each participant passed the test of selective influence, the Houpt-Townsend statistic (D+, D-), the mean interaction contrast (MIC), and the identified processing model. Bold D+ and D- statistics indicate a significant Houpt-Townsend statistic at $p < 0.33$.

uli generalize in a dynamic environment. A dynamic environment includes highly correlated movement of objects across time. If movement across time is redundant across various types of multispectral imagery displayed simultaneously, they may provide additional speed-ups that are not provided by a single, composite image. Alternatively, each multispectral image may provide different information about the dynamic movement in the scene leading to more efficient processing when the images are confined in a spatially overlapping reference plain. SFT offers a way to compare these techniques and can illustrate why or where strategies may inhibit the processing of each contributing multispectral image.

Methods

The video fusion stimuli are from OTCBVS Benchmark Dataset (`http://vcipl-okstate.org/pbvs/bench/Data/03/download.html`). From the video clips we selected 109 300ms video clips . Selection was based on at least one person (target) walking leftward or rightward (67 walking right, 42 walking left). We required that, if multiple people were in view in a clip, they were all walking in the same direction. On any given trial, the video clip may be vertically flipped, thereby creating a total of 218 possible stimuli (109 walking right, 109 walking left). All stimuli had a resolution of 224×224. Participants were required to respond using corresponding mouse buttons, left-click for leftward movement and right-click for rightward movement. The following results are for 5 participants who volunteered their time to participate.

DISTRIBUTION A: Distribution approved for public release.

Figure 36: An example frame of a typical dynamic stimulus illustrating a LWIR-only trial (left), a visible-only trial (center), or an algorithmically combined video (right).

Table 15: Number of trials to compute the algorithmic and cognitive capacity with dynamic multi-sensor imagery.

	visible-only	LWIR-only	Multi-sensor	Total
Cognitive	120	120	120	380
Algorithmic	120	120	120	380

The first experiment focused on calculating the capacity coefficient for algorithm-fused and cognitive-fused dynamic imagery (Table 15). There were 15 trials of each visible-only, LWIR-only, and multi-sensor for both algorithm-fused and cognitive-fused imagery (90 trials total; 3min, 45sec) as practice at the beginning of the session. An example stimulus is shown in Figure 36.

Note that visible-only and LWIR-only differ in where they will be presented on the screen depending on the type multi-sensor fusion for the respective block of trials: when the individual sensor trials are interleaved with algorithmic fusion they will be presented in the center of the screen, when the individual sensor trials are interleaved with multi-sensor cognitive fusion they will be presented either on the left (LWIR) or right (visible) of the center screen. We stress a distinction in the two conditions to eliminate the potential for increases in visual attention demand when comparing multi-sensor performance to our predicted baseline estimated from individual sensor performance.

Participants needed 80% accuracy in the practice before continuing to experimental trials. If participants were below 80% accuracy, the practice trials were repeated. The total length of the first experiment was one 1-hour session with approximately 37 mins, 25secs of total experiment run time. The completion time of the session includes training trials (assuming the participant passes training requirements), the cognitive fusion condition, and the algorithmic fusion condition. Participants had the opportunity to take a break between each block of trials.

In summary, RTs and accuracy with the fused imagery were worse than single-sensor images (Figure 37). Responses for the LWIR only were faster and

DISTRIBUTION A: Distribution approved for public release.

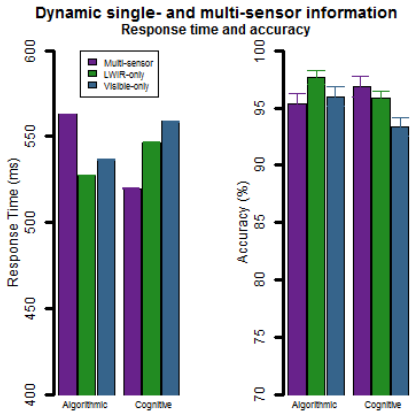

Figure 37: Mean correct RTs (left) and accuracy (right) for each sensor type for each fusion method in the dynamic video stimuli. Cognitive fusion (visible and LWIR alone randomly presented on the left/right of center) and algorithmic fusion (visible and LWIR alone presented in the center of the screen). Error bars represent the standard error of the mean (Jarmasz & Hollands, 2009).

more accurate than visible only. Likewise for algorithmic fusion, visible and LWIR-only were consistently faster but less accurate than algorithmic fusion. Interestingly, the reverse was shown for cognitive fusion. Mean response times were slower and less accurate for single-sensor trials than cognitive fusion trials (sensors presented side-by-side). Mean response times for cognitive fusion were significantly faster than algorithmic fusion ($t(1150) = 5.07, p < .05$).

Further individual-level analyses of the capacity coefficient allows us to examine how the cognitive processing changes across the manipulated fusion type and sensor condition by participant. Separate analyses were conducted for algorithmic and cognitive fusion for the dynamic fusion stimuli. Individual scores are reported in Table 16. The capacity coefficient was below 1 (i.e., limited capacity) for some time for algorithmic fusion for all participants. Individuals capacity z-scores in the algorithmic fusion condition ranged from -9.9 to -6.2. Four participants had performance patterns indicated a capacity coefficient below 1 for some time with capacity z-scores ranging from -6.11 to -2.25. One participant exhibited an workload capacity that was not significantly limited, with a z-score of -1.85. At the group level, both algorithmic and cognitive fusion were significantly less than one, i.e., limited capacity (Algorithmic: $t(4) = -13.54, p < .05$, Cognitive: $t(4) = -4.78, p < .05$). Algorithmic fusion was significantly more limited than cognitive fusion, $t(4) = -6.32, p < .05$.

DISTRIBUTION A: Distribution approved for public release.

Table 16: Individual level capacity, z-score, and statistical significance for algorithmic and cognitive fusion of dynamic multi-sensor images compared to UCIP model in a orientation of walking discrimination task.

	Algorithmic		Cognitive	
Subject	Capacity	z-score	Capacity	z-score
1	Limited	-8.24***	Limited	-6.11***
2	Limited	-8.02***	Limited	-2.25*
3	Limited	-6.18***	Unlimited	-1.85
4	Limited	-9.93***	Limited	-3.88***
5	Limited	-7.93***	Limited	-3.95***

References

Ahumada, A. J. (2002). Classification image weights and internal noise level estimation. *Journal of Vision*, *2*, 121-131.

Ahumada, A. J., & Krebs, W. K. (2000). Signal detection in fixed pattern chromatic noise. *Investigative Ophthalmology and Visual Science*, *41*, 3796-3804.

Ahumada, A. J., & Lovell, J. (1971). Stimulus features in signal detection. *Journal of the Acoustic Society of America*, *49*, 1751-1756.

Altieri, N., & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology*, *2*, 1-15.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*, 154-179.

Barlow, H. B., & Reeves, B. C. (1979). The versatility and absolute efficiency of detecting mirror symmetry in random dot displays. *Vision Research*, *19*, 783–793.

Békésy, G., & Wever, E. G. (1960). *Experiments in hearing* (Vol. 8). McGraw-Hill New York.

Bittner, J. (2015). Areas of visual information utilized by humans in multispectral fused imagery using classification images. 56th Annual Meeting of the Psychonomics Society; Chicago, IL. (poster)

Blasch, E., & Plano, S. (2005). Proactive decision fusion for site security. *Information Fusion*, *2*, 1-8.

Blum, R. (2006). *Multi-sensor image fusion and its applications*. Boca Raton, FL: Taylor and Francis.

Brainard, D. H., Williams, D. R., & Hofer, H. (2008). Trichromatic reconstruction from the interleaved cone mosaic: Bayesian model and the color appearance of small spots. *Journal of Vision*, *8*(5), 15.

Burns, D. M., Pei, L., Houpt, J. W., & Townsend, J. T. (2009). *Facial perception as a configural process.* Poster presented at: Annual Meeting of the Cognitive Science Society.

Burt, P. J., & Adelson, E. H. (1983). The laplacian pyramid as a compact

DISTRIBUTION A: Distribution approved for public release.

image code. *Communications, IEEE Transactions*, *31*, 532-540.
Burt, P. J., & Kolczynski, R. J. (1993). Enhanced image capture through fusion. In (p. 173-182). IEEE Fourth International Conference.
Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.
Dixon, T. D., Canga, E. F., Noyes, J. M., Troscianko, T., Nikolov, S. G., Bull, D. R., & Canagarajah, C. N. (2006). Methods for the assessment of fused images. *ACM Transactions on Applied Perception (TAP)*, *3*, 309-332.
Dixon, T. D., Li, J., Noyes, J. M., Troscianko, T., Nikolov, S. G., Lewis, J. J., & Canagarajah, C. N. (2007). Scanpath assessment of visible and infrared side-by-side and fused video displays. In *Proceedings of the 10th international conference on information fusion* (p. 1-8). Canada.
Dong, J., Zhuang, D., Huang, Y., & Fu, J. (2009). Advances in multi-sensor data fusion: Algorithms and applications. *Sensors*, *9*, 7771-7784.
Donkin, C., Little, D. R., & Houpt, J. W. (2014). Assessing the speed-accuracy trade-off effect on the capacity of information processing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(3), 1-20.
Donnelly, N., Cornes, K., & Menneer, T. (2012). An examination of the processing capacity of features in the Thatcher illusion. *Attention, Perception & Psychophysics*, *74*, 1475-1487.
Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid monte carlo. *Physics letters B*, *195*(2), 216–222.
Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, *87*, 272-300.
Dzhafarov, E. N. (2003). Selective influence through conditional independence. *Psychometrika*, *68*, 7-25.
Dzhafarov, E. N., Schweickert, R., & Sung, K. (2004). Mental architectures with selectively influenced but stochastically interdependent components. *Journal of Mathematical Psychology*, *48*, 51-64.
Eidels, A., Houpt, J. W., Pei, L., Altieri, N., & Townsend, J. T. (2011a). Nice guys finish fast, bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology*, *55*, 176-190.
Eidels, A., Houpt, J. W., Pei, L., Altieri, N., & Townsend, J. T. (2011b). Nice guys finish fast, bad guys finish last: Facilitatory vs. inhibitory interaction in parallel systems. *Journal of Mathematical Psychology*, *55*, 176-190.
Eidels, A., Townsend, J. T., Hughes, H. C., & Perry, L. A. (2015). Evaluating perceptual integration: uniting response-time-and accuracy-based methodologies. *Attention, Perception, & Psychophysics*, *77*(2), 659–680.
Essock, E., Sinai, M., McCarley, J., Krebs, W. K., & DeFord, J. (1999). Perceptual ability with real-world nighttime scenes: image-intensified, infrared, and fused-color imagery. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *41*, 438-452.
Ferguson, T. S. (1973). A Bayesian anlaysis of some nonparametric problems. *Annals of Statistics*, *1*, 209-230.
Fifić, M., Nosofsky, R. M., & Townsend, J. T. (2008). Information-processing architectures in multidimensional classification: A validation test of the systems factorial technology. *Journal of Experimental Psychology: Human*

DISTRIBUTION A: Distribution approved for public release.

Perception and Performance, *34*, 356-375.

Fifić, M., & Townsend, J. T. (2010). Information-processing alternatives to holistic perception: Identifying the mechanisms of secondary-level holism within a categorization paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1290-1313.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, *47*(6), 381.

Fox, L., Elizabeth, & Houpt, J. W. (2016). The perceptual processing of fused multispectral imagery. *Cognitive Research: Principles and Implications*. (in press)

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, *32*, 675–701.

Gaspar, C. M., Bennett, P. J., & Sekuler, A. B. (2008). The effects of face inversion and contrast-reversal on efficiency and internal noise. *Vision Research*, *48*(8), 1084–1095.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision research*, *51*(7), 771–781.

Glasgow, R. L., Marasco, P. L., Havig, P. R., Martinsen, G. L., Reis, G. A., & Heft, E. L. (2003). Psychophysical measurement of night vision goggle noise. In *Aerosense 2003* (p. 164-173).

Gold, J. M., Bennett, P. J., & Sekuler, A. B. (1999). Identification of band-pass filtered faces and letters by human and ideal observers. *Vision Research*, *39*, 3537–3560.

Gold, J. M., Sekuler, A. B., & Bennett, P. J. (2004). Characterizing perceptual learning with external noise. *Cognitive Science*, *28*, 167-207.

Gold, J. M., Tadin, D., Cook, S. C., & Blake, R. (2008). The efficiency of biological motion perception. *Perception & Psychophysics*, *70*(1), 88–95.

Hall, D. L., & Llinas, J. (1997). An introduction to multisensor data fusion. In (Vol. 85, p. 6-23). IEEE.

Hall, D. L., & Steinberg, A. (2000). Dirty secrets in multisensor data fusion. San Antonio, TX: National Symposium on Sensor Data Fusion (NSSDF).

Harris, J. (2008). Mindmodeling@home: a large-scale computational cognitive modeling infrastructure. In *Proceedings of the sixth annual conference on systems engineering research 2008* (p. 246-252). Los Angeles, CA, USA.

Heathcote, A. (2004). Fitting wald and ex-wald distributions to response time data: An example using functions for the s-plus package. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 678–694.

Heathcote, A., Brown, S., Wagenmakers, E. J., & Eidels, A. (2010). Distribution-free tests of stochastic dominance for small samples. *Journal of Mathematical Psychology*, *54*, 454-463.

Houpt, J. W., & Bittner, J. L. (n.d.). *Analyzing thresholds and efficiencies with hierarchical Bayesian linear regression.* (in prep)

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2013). Systems Factorial Technology with R. *Behavior Research Methods*,

DISTRIBUTION A: Distribution approved for public release.

46, 307-330.

Houpt, J. W., Blaha, L. M., McIntire, J. P., Havig, P. R., & Townsend, J. T. (2014). Systems factorial technology with r. *Behavior Research Methods*, *46*, 307-330.

Houpt, J. W., & Burns, D. M. (2016). Statistical analyses for systems factorial technology. In D. R. Little, N. Altieri, M. Fifić, & C.-T. Yang (Eds.), *Systems factorial technology: A theory driven methodology for the identification of perceptual and cognitive mechanisms.* Elsevier.

Houpt, J. W., & Fifić, M. (2013). A hierarchical approach to distinguishing serial and parallel processing. In *Annual meeting of the psychonomic society.* Toronto, ON.

Houpt, J. W., Heathcote, A., & Eidels, A. (in press). Bayesian analyses of cognitive architecture. *Psychological Methods*.

Houpt, J. W., Heathcote, A., Eidels, A., Medeiros-Ward, N., Watson, J., & Strayer, D. (2012). Capacity coefficient variations. Psychonomics Society Meeting. (Poster)

Houpt, J. W., MacEachern, S. N., Peruggia, M., Townsend, J. T., & Van Zandt, T. (2016). Semiparametric Bayesian approaches to systems factorial technology. *Journal of Mathematical Psychology*. Retrieved from `http://www.sciencedirect.com/science/article/pii/S0022249616000249` doi: http://dx.doi.org/10.1016/j.jmp.2016.02.008

Houpt, J. W., Townsend, J., & Donkin, C. (2014). A new perspective on visual word processing efficiency. *Acta Psychologica*, *145*, 118-127.

Houpt, J. W., & Townsend, J. T. (2010a). The statistical properties of the survivor interaction contrast. *Journal of Mathematical Psychology*, *54*, 446-453.

Houpt, J. W., & Townsend, J. T. (2010b). The statistical properties of the survivor interaction contrast. *Journal of Mathematical Psychology*, *54*, 446-453.

Houpt, J. W., & Townsend, J. T. (2011). An extension of SIC predictions to the Wiener coactive model. *Journal of Mathematical Psychology*, *55*, 267-270.

Houpt, J. W., & Townsend, J. T. (2012). Statistical measures for workload capacity analysis. *Journal of Mathematical Psychology*, *56*, 341-355.

Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2005). *Bayesian survival analysis.* Wiley Online Library.

Irwin, D. E. (1991). Information integration across saccadic eye movements. *Cognitive Psychology*, *23*, 420-456.

Jarmasz, J., & Hollands, J. G. (2009). Confidence intervals in repeated-measures designs: The number of observations principle. *Canadian Journal of Experimental Psychology*, *63*, 124-138.

Johnson, S. A., Blaha, L. M., Houpt, J. W., & Townsend, J. T. (2010). Systems factorial technology provides new insights on global–local information processing in autism spectrum disorders. *Journal of Mathematical Psychology*, *54*, 53-72.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-

DISTRIBUTION A: Distribution approved for public release.

Hall.
Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.
Kirk, R. E. (2012). *Experimental design: Procedures for the behavioral sciences* (4th ed.). Thousand Oaks, CA: Sage.
Klein, G., Moon, B. M., & Hoffman, R. R. (2006). Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems*, *21*, 88-92.
Klugkist, I., Kato, B., & Hoijtink, H. (2005). Bayesian model selection using encompassing priors. *Statistica Neerlandica*, *59*, 57-69.
Kontsevich, L. L., & Tyler, C. W. (1999). Bayesian adaptive estimation of psychometric slope and threshold. *Vision research*, *39*(16), 2729–2737.
Krebs, W. K., McCarley, J. S., Kozek, T., Miller, G. M., Sinai, M. J., & Werblin, F. S. (1999). An evaluation of a sensor fusion system to improve drivers' nighttime detection of road hazards. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *43*, 1333-1337.
Krebs, W. K., Scribner, D. A., Miller, G. M., Ogawa, J. S., & Schuler, J. (1998). Beyond third generation: A sensor fusion targeting flir pod for the f/a-18. In B. Dasarathy (Ed.), (Vol. 3376, p. 129-140). Bellingham, WA: SPIE - International Society for Optical Engineering.
Krebs, W. K., & Sinai, M. J. (2002). Psychophysical assessments of image-sensor fused imagery. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *44*, 257-271.
Krishnamoorthy, S., & Soman, K. P. (2010). Implementation and comparative study of image fusion algorithms. *International Journal of Computer Applications*, *9*, 25-35.
Kruschke, J. K. (2010). *Doing bayesian data analysis: A tutorial with R and BUGS*. Amsterdam: Academic Press.
Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science*, *6*, 299-312.
Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*(5), 8. doi: 10.1167/5.5.8
Lawrence, M. A. (2012). ez: Easy analysis and visualization of factorial experiments. [Computer software manual]. Retrieved from `http://CRAN.R-project.org/package=ez` (R package version 4.1-1)
Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1-27.
Liu, Z., & Kersten, D. (2003). Three-dimensional symetric shapes are discriminated more efficiently tahn asymmetric ones. *Journal for the Optical Society of America A*, *20*(7), 1331-1340. doi: 10.1167/5.5.8
Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide (second edition)*. Mahwah, NJ: Erlbaum.
McCarley, J., & Krebs, W. K. (2000). Visibility of road hazards in thermal, visible, and sensor-fused nighttime imagery. *Applied Ergonomics*, *31*, 523-530.

DISTRIBUTION A: Distribution approved for public release.

McCarley, J., & Krebs, W. K. (2006). The psychophysics of sensor fusion: A multidimensional signal detection analysis. In (Vol. 50, p. 2094-2098).

Microsoft Azure. (n.d.). `http://azure.microsoft.com`.

Miller, J. (1982). Divided attention: Evidence for coactivation with redundant signals. *Cognitive Psychology*, *14*, 247-279.

Mordkoff, J. T., & Yantis, S. (1991a). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*, 520-538.

Mordkoff, J. T., & Yantis, S. (1991b). An interactive race model of divided attention. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(2), 520.

Morrison, D. J., & Schyns, P. G. (2001). Usage of spatial scales for the categorization of faces, objects, and scenes. *Psychonomic Bulletin and Review*, *8*, 454-469.

Neriani, K. E., Pinkus, A. R., & Dommett, D. W. (2008). An investigation of image fusion algorithms using a visual performance-based image evaluation methodology. In (Vol. 6968). In SPIE Defense and Security Symposium. International Society for Optics and Photonics.

Ohio Supercomputer Center. (1987). *Ohio supercomputer center.* `http://osc.edu/ark:/19495/f5s1ph73`.

Ohio Supercomputer Center. (2012). *Oakley supercomputer.* `http://osc.edu/ark:/19495/hpc0cvqn`.

Peirce, J. (2009). Generating stimuli for neuroscience using psychopy. *Front. Neuroinform. 2:10*. doi: 10.3389/neuro.11.010.2008

Petrović, V. (2007). Subjective tests for image fusion evaluation and objective metric validation. *Information Fusion*, *8*, 208-216.

Petrović, V., & Xydeas, C. (2004). Evaluation of image fusion performance with visible differences. *Computer Vision-ECCV*, 380-391.

Piella, G., & Heijmans, H. (2003). A new quality metric for image fusion. In (Vol. 3, p. 111-173). ICIP International Conference: Image Processing.

Pollatsek, A., Rayner, K., & Collins, W. E. (1984). Integrating pictorial information across eye movements. *Journal of Experimental Psychology: General*, *113*, 426-442.

Qu, G., Zhang, D., & Yan, P. (2002). Information measure for performance of image fusion. *Electronics Letters*, *38*, 313-315.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org` (ISBN 3-900051-07-0)

Raab, D. (1962). Statistical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences*, *24*, 574-590.

Rayner, K., McConkie, G. W., & Zola, D. (1980). Integrating information across eye movements. *Cognitive pscyhology*, *12*, 206-226.

Reis, G. A., Marasco, P. L., Havig, P. R., & Heft, E. L. (2004). Psychophysical measurement of night vision goggle noise using a binocular display. In *Defense and security* (p. 13-24).

Repperger, D. W., Havig, P. R., Reis, G. A., Farris, K. A., McIntire, J. P.,

DISTRIBUTION A: Distribution approved for public release.

Townsend, J. T., ... Houpt, J. W. (2009). Studies on hazard functions and human performance. *The Ohio Journal of Science*, *109*.

Rouder, J. N., & Lu, J. (2005). An introduction to bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*(4), 573–604.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Bayes factors for ANOVA designs. *Journal of Mathematical Psychology*, *56*, 356-374.

Rousselet, G., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature neuroscience*, *5*, 629-630.

Ryan, D. M., & Tinkler, R. D. (1995). Night pilotage assessment of image fusion. In (p. 50-67). International Society for Optics and Photonics: SPIEs Symposium on OE/Aerospace Sensing and Dual Use Photonics.

Scharff, A., Palmer, J., & Moore, C. M. (2011). Evidence of fixed capacity in visual object categorization. *Psychonomic Bulletin Review*, *18*, 713-721.

Schreiber, B. T., Stock, W. A., & Bennett, W., Jr. (2006). *Distributed mission operations within-simulator training effectiveness baseline study: Metric development and objectively quantifying the degree of learning* (Tech. Rep.). Mesa, AZ: Air Force Research Laboratory. Retrieved from `http://www.dtic.mil/dtic/tr/fulltext/u2/a461867.pdf`

Schwarz, W. (1989). A new model to explain the redundant-signals effect. *Perception & Psychophysics*, *46*, 498-500.

Schwarz, W. (1994). Diffusion, superposition, and the redundant-targets effect. *Journal of Mathematical Psychology*, *38*, 504-520.

Sinai, M. J., McCarley, J. S., & Krebs, W. K. (1999). Scene recognition with infrared, low-light, and sensor-fused imagery. In (p. 1-9). Ann Arbor, MI: IRIS Specialty Groups on Passive Sensors IRIA.

Smeelen, M. A., Schwering, P. B., Toet, A., & Loog, M. (2014). Semi-hidden target recognition in gated viewer images fused with thermal ir images. *Information Fusion*, *18*, 131-147.

Stan Development Team. (2014a). *Stan: A c++ library for probability and sampling, version 2.5.0.* Retrieved from `http://mc-stan.org/`

Stan Development Team. (2014b). Stan modeling language users guide and reference manual, version 2.5.0 [Computer software manual]. Retrieved from `http://mc-stan.org/`

Steele, P., & Perconti, P. (1997). Part-task investigation of multispectral image fusion using gray scale and synthetic color night vision sensor imagery for helicopter pilotage. In W. R. . D. Clement (Ed.), (Vol. 3062, p. 88-100). Bellingham, WA: SPIE -Aerospace/Defense Sensing, Simulation and Controls.

Tanner Jr, W. P., & Birdsall, T. (1958). Definitions of d? and η as psychophysical measures. *The Journal of the Acoustical society of America*, *30*, 922.

Thiele, J. E., & Rouder, J. N. (2016). *Bayesian analysis for systems factorial technology.* Retrieved from `http://pcl.missouri.edu/sites/default/`

DISTRIBUTION A: Distribution approved for public release.

files/p_8.pdf
Tjan, B. S., Braje, W. L., Legge, G. E., & Kersten, D. (1995). Human efficiency for recognizing 3-d objects in luminance noise. *Vision research*, *35*(21), 3053–3069.
Toet, A. (2013). Registration of a dynamic multimodal target image test set for the evaluation of image fusion techniques. *The Air Force Office of Scientific Research and European Office of Aerospace Research and Development*. (Grant No.: FA8655-11-1-3015. Soesterberg, The Netherlands.)
Toet, A., & Franken, E. M. (2003). Perceptual evaluation of different image fusion schemes. *Displays*, *24*, 25-37.
Toet, A., & Hogervorst, M. A. (2009). Triclobs portable triband color lowlight observation system. In (Vol. 7345, p. 1-11). SPIE Defense, Security, and Sensing.
Toet, A., & Hogervorst, M. A. (2012). Progress in color night vision. *Optical Engineering*, *51*, 1-19.
Toet, A., Hogervorst, M. A., Nikolov, S. G., Lewis, J. J., Dixon, T. D., Bull, D. R., & Canagarajah, C. N. (2010a). Towards cognitive image fusion. *Information Fusion*, *11*, 95-113.
Toet, A., Hogervorst, M. A., Nikolov, S. G., Lewis, J. J., Dixon, T. D., Bull, D. R., & Canagarajah, C. N. (2010b). Towards cognitive image fusion. *Information Fusion*, *11*, 95-113.
Toet, A., Ljspeert, I., Waxman, A., & Aguilar, M. (1997). Fusion of visible and thermal imagery improves situational awareness. In J. G. Verly (Ed.), (Vol. 3088, p. 177-188). Bellingham, WA: SPIE - Enhanced and Synthetic Vision.
Townsend, J. T. (1974). Issues and models concerning the processing of a finite number of inputs. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (p. 133-168). Hillsdale, NJ: Erlbaum Press.
Townsend, J. T., & Altieri, N. (2012). An accuracy-response time capacity assessment function that measures performance against standard parallel predictions. *Psychological Review*, *119*, 500-516.
Townsend, J. T., & Ashby, F. G. (1983a). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.
Townsend, J. T., & Ashby, F. G. (1983b). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.
Townsend, J. T., & Fifić, M. (2004). Parallel and serial processing and individual differences in high-speed scanning in human memory. *Perception & Psychophysics*, *66*, 953-962.
Townsend, J. T., & Nozawa, G. (1995a). Spatio-temporal properties of elementary perception: An investigation of parallel, serial and coactive theories. *Journal of Mathematical Psychology*, *39*, 321-360.
Townsend, J. T., & Nozawa, G. (1995b). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and coactive theories. *Journal of Mathematical Psychology*, *39*, 321-359.
Townsend, J. T., & Wenger, M. J. (2004a). A theory of interactive parallel

DISTRIBUTION A: Distribution approved for public release.

processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003-1035.

Townsend, J. T., & Wenger, M. J. (2004b). A thoery of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review*, *111*, 1003- 1035.

Urban, F. M. (1910). The method of constant stimuli and its generalizations. *Psychological Review*, *17*(4), 229.

Van Zandt, T. (2002). Analysis of response time distributions. In J. T. Wixted & H. Pashler (Eds.), *Stevens' handbook of experimental psychology* (3rd ed., Vol. 4, p. 461-516). New York: Wiley Press.

Watson, A., & Pelli, D. (1983). Quest: A bayesian adaptive psychometric method. *Perception and Psychophysics*, *33*, 13-120.

Watson, A. B., & Pelli, D. G. (1983). QUEST: A bayesian adaptive psychometric method. *Perception & Psychophysics*, *33*(2), 113–120.

Xydeas, C. S., & Petrović, V. S. (2000). Objective pixel-level image fusion performance measure. In (p. 89-98). International Society for Optics and Photonics: AeroSense 2000.

Yang, B., Jing, Z. L., & Zhao, H. T. (2010). Review of pixel-level image fusion. *Journal of Shanghai Jiaotong University (Science)*, *15*, 6-12.

Yang, C.-T. (2011). Relative saliency in change signal affects perceptual comparison and decision processes in change detection. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1708-1728.

Yang, C.-T., Chang, T.-Y., & Wu, C.-J. (2012). Relative change probability affects the decision process of detecting multiple feature changes. *Journal of Experimental Psychology: Human Perception and Performance*. (Advance online publication)

Yang, C.-T., Hsu, Y.-F., Huang, H.-Y., & Yeh, Y.-Y. (2011). Relative salience affects the process of detecting changes in orientation and luminance. *Acta Psychologica*, *138*, 377-389.

Yang, C.-T., Little, D. R., & Hsu, C.-C. (2014). The influence of cueing on attentional focus in perceptual decision making. *Attention, Perception, & Psychophysics*, *76*(8), 2256–2275.

Yong, Z., Weiqi, J., & Rui, X. (2010). Assessment method to fusion effect based on structural similarity comparison in fusion images. In (Vol. 7820). In International Conference on Image Processing and Pattern Recognition in Industrial Engineering: International Society for Optics and Photonics.

Zhang, H., & Houpt, J. W. (2016). Assessing multispectral image fusion with systems factorial technology. In *Proceedings of the 60th annual meeting of the human factors and ergonomics society.* Sage.

DISTRIBUTION A: Distribution approved for public release.

AFOSR Deliverables Submission Survey

Response ID:7300 Data

1.

Report Type

Final Report

Primary Contact Email

Contact email if there is a problem with the report.

joseph.houpt@wright.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

8122022509

Organization / Institution name

Wright State University

Grant/Contract Title

The full title of the funded effort.

Dynamic Generalizations of Systems Factorial Technology for Modeling Perception of Fused Information

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-13-1-0087

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Joseph W. Houpt

Program Officer

The AFOSR Program Officer currently assigned to the award

James Lawton

Reporting Period Start Date

03/01/2013

Reporting Period End Date

08/31/2016

Abstract

Models are a fundamental part of understanding cognition. The advantages of cognitive modeling are particularly clear when attempting to understand how changes in a cognitive task lead to changes in performance. Systems factorial technology (SFT) can be used to explain and understand why there are differences in performance, not just that there is a difference. In this project, we have extended the applicability of SFT to more complex environments than the basic perceptual experiments to which it has been previously applied. This included extensions of the statistical analyses to include hierarchical parametric Bayesian modeling and semi- and non-parametric modeling. We then applied SFT in both basic visual search studies and in task requiring the use of multispectral imagery.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

DISTRIBUTION A: Distribution approved for public release.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

sf0298.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

afosr-final-report(1).pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

Accepted publications since last report:

Houpt, J. W. and Burns, D. M. (in press). Statistical analyses for Systems Factorial Technology. In D. R. Little, N. Altieri, M. Fific, & C-T. Yang (Eds.), Systems Factorial Technology: A Theory Driven Methodology for the Identification of Perceptual and Cognitive Mechanisms. Elsevier.

Fox, E. L. and Houpt, J. W. (in press). The perceptual processing of fused multispectral imagery. Cognitive Research: Principles and Implications.

Houpt, J. W., Heathcote, A. and Eidels, A. (in press). Bayesian analyses of cognitive architecture. Psychological Methods.

Hammack, T., Cooper, J., Flach, J., and Houpt, J. W. (in press). Toward an ecological theory of rationality: Debunking the hot hand `illusion.' Ecological
Psychology.

New discoveries, inventions, or patent disclosures:
Do you have any discoveries, inventions, or patent disclosures to report for this period?

No

Please describe and include any notable dates

Do you plan to pursue a claim for personal or organizational intellectual property?

Changes in research objectives (if any):

Change in AFOSR Program Officer, if any:

Extensions granted or milestones slipped, if any:

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

DISTRIBUTION A: Distribution approved for public release.

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Nov 30, 2016 21:51:08 Success: Email Sent to: joseph.houpt@wright.edu

DISTRIBUTION A: Distribution approved for public release.